2000

A Construct Validity Study of the New Jersey State Assessment Program : Grade 11 High School Proficiency Test (HSPT11), Early Warning Test (EWT), Grade Eight Proficiency Assessment (GEPA) and Elementary School Proficiency Assessment (ESPA)

Maureen O'Sullivan Lally
Seton Hall University

A CONSTRUCT VALIDITY STUDY OF THE NEW JERSEY STATE ASSESSMENT PROGRAM - GRADE 11 HIGH SCHOOL PROFICIENCY TEST (HSPT11), EARLY WARNING TEST (EWT), GRADE EIGHT PROFICIENCY ASSESSMENT (GEPA) AND ELEMENTARY SCHOOL PROFICIENCY ASSESSMENT (ESPA)

2000

MAUREEN O' SULLIVAN LALLY

A CONSTRUCT VALIDITY STUDY OF THE NEW JERSEY STATE ASSESSMENT PROGRAM - GRADE 11 HIGH SCHOOL PROFICIENCY TEST (HSPT11), EARLY WARNING TEST (EWT), GRADE EIGHT PROFICIENCY ASSESSMENT (GEPA) AND ELEMENTARY SCHOOL PROFICIENCY ASSESSMENT (ESPA)

BY

MAUREEN O' SULLIVAN LALLY

Dissertation Committee

Elaine Walker, Ph.D., Mentor
Reverend Kevin Hanbury, Ed.D.
Margaret M. McCluskey, Ed.D.
Anne M. Wilkins, Ed.D.

Submitted in partial fulfillment of the
requirements of the Degree of Doctor of Education
Seton Hall University

2000

NOTICE OF COPYRIGHT

ACKNOWLEDGEMENTS

I would like to extend my appreciation to all those who provided time, encouragement and their experience as I pursued and completed this important goal.

I was fortunate to have committee members who were supportive and pragmatic while encouraging a timely completion. Thank you to my mentor, Dr. Elaine Walker, and committee members and friends, Reverend Kevin Hanbury, Dr. Margaret McCluskey and Dr. Anne Wilkins. Thank you, also, to Elena Martino who patiently provided word processing and formating guidance.

For the many years that I talked about this goal, and the final year when I became completely absorbed in the process, I would like to thank my family for their support and the willingness to move with agility around the omnipresent computer and boxes of research. I am particularly grateful that Jerry shouldered the tedious task of matching twenty-nine pages of references to the appropriate citations.

Finally, I would like to thank my parents who provided a vision for this goal. From the time my father received his doctorate from Seton Hall, my mother became convinced that I should do the same. Many years later, that goal has been reached.

TABLE OF CONTENTS

LIST OF TABLES

Chapter I

INTRODUCTION

In 1122 BC, the Chow Dynasty instituted biannual pupil exams (Nitko, 1983). Since that historic event, educators and tests have had an uneasy relationship. At times assessments have been embraced; at other times, eyed with scorn. The current emphasis on accountability has brought renewed attention to the topic.

For much of assessment history, tests were used to select and sort. An early leader in this movement was the British educational system that used national exams to separate students into different educational tracks (Lemann, 1999; IRA, 1999). While educators in the United States have not been immune to this age-old practice of dividing by ability, a second use for tests developed during the accountability movement of the 1960's. As related by Smith (1986), conducting long-term studies to determine the success of graduates was considered to be prohibitively expensive. On the other hand, measuring the performance of students while they were still in school was straightforward and economically practical. "The technology of educational testing could readily be applied in evaluating the productivity of school in terms of scholastic accomplishment. Teachers were already equipped to carry out the data collection.

Student achievement could be sampled and assessed, and school performance evaluated" (p. 166).

This mindset was further supported by national legislation. Title 1 of the Elementary and Secondary Education Act (ESEA), first passed in 1965, included evaluation as an essential part of the effort to improve conditions of the poor and to produce curriculum that responded to changing economic conditions. ESEA was, and is, the main federal law impacting K-12 schools. Through Title 1, districts began to consider the constant monitoring of student progress as a measure of the success of teachers and schools.

As the United States entered the last decades of the Twentieth Century, the monitoring supported by ESEA led to increased frustration with a perceived lack of student achievement. In 1983, the National Commission on Excellence, recommended "replacing the unfocused school experience that was prevalent during the late 1960s and '70s with meaningful curriculum standards and assessments" (Hespe, 1999d).

In 1989, the historic national educational summit in Charlottesville, Virginia, echoed this vision, proposing national goals. In 1996, the second national educational summit, held in Palisades, New York, moved to the next level by focusing on the need to set standards and create aligned assessments.

Then, on September 30 and October 1, 1999, twenty-three governors along with ninety-one corporate executives and education leaders gathered at the third national education summit, again in Palisades. As reported by Olson and Hoff (1999), Patton and Thompson (1999), the conference supported the work of the second summit and adopted an action statement that addressed strengthening accountability.

President Clinton proved to be a supporter of the accountability philosophy. In 1994, the president made his position known with the reauthorization of ESEA. This reauthorization required states and districts receiving federal dollars to set up systems of standards and aligned assessments. Clinton's May 1999 ESEA plan continued this concept of accountability, requiring performance report cards at the state, district and school levels (Robelen, 1999).

States listened to the call from President Clinton and the national summits. According to the Mid-continent Regional Educational Laboratory (McRel), forty-nine states accepted the accountability challenge and developed standards documents. As for assessment, Jerald, in Quality Counts 2000, reported that forty-eight states and the District of Columbia administered testing programs during the 1999-2000 school year. While this number included neither Iowa nor Nebraska, virtually every

public-school student in Iowa took the same standardized Iowa Test of Basic Skills (ITBS).

Of the forty-eight states with assessment programs, twenty-four have high-stakes tests that high school students must pass in order to get a diploma. This follows the National Association of State Boards of Education October 1997 recommendation that called for attaching high stakes, such as high school graduation and grade promotion, to students' performance on state assessments (Manzo, 1997).

It is obvious that high standards, high-stakes tests, and accountability are the current trends in public education. During 2000, forty states plan to issue report cards on schools based on their test performance, twenty-one plan to issue overall ratings for schools based on performance, and eighteen expect to have the legal authority to close, take over, or move staffs of failing schools (Jerald, 2000). As reported by the Associated Press on June 8, 2000, both presidential candidates, Republican George W. Bush and Democrat Al Gore, have proposed tying federal education dollars to states' test scores. In the March 1999 issue of Educational Leadership, McRel's Schmoker and Marzano wrote, "State and standardized assessments do not measure everything we deem important, but success on such tests

in this age of accountability is vital. Strong standardized scores earn us the trust of our communities..." (p. 20).

Dr. Diane Ravitch seconded this opinion. A research professor at the New York University School of Education, senior fellow at the Brookings Institution and a former assistant U.S. education secretary, Dr. Ravitch presented the keynote address at the October 1999 annual convention of the New Jersey School Boards Association (NJSBA). In that address she observed that the revolution of rising expectations was an accurate reflection of economic and social realities.

Not everyone agreed with this emphasis on assessment. In an unusual and dramatic move, the Board of Directors of the International Reading Association (IRA) adopted a May 1999 position statement opposed to high-stakes testing. This statement reflected earlier critics such as W. James Popham and Linda Darling-Hammond. An expert on educational testing and professor emeritus of the University of California, Los Angeles, Dr. Popham wrote: "In an evidence-oriented enterprise, those who control the evidence-gathering mechanisms control the entire enterprise. Control is achieved because teachers have little choice but to teach to the tests" (Smith, 1986, p. 130). Linda Darling-Hammond's Rand Corporation study concurred: (when) "important decisions are based on test scores...teachers are

more likely to teach to the tests and less likely to bother with nontested activities" (Smith, 1986, p. 134). According to Lemann (1999), Bill Trumbell, second president of the Educational Testing Service (ETS), used the Procrustean bed, a metaphor of Greek mythology, for the same concept. Procrustes, an Attic villian, would waylay passerby and contort them into conformance with his misshapen, uncomfortable bed. It would appear that, in the minds of some, today's state assessments have become the modern Procrustean bed.

While debate surrounds the issues of standards, assessments and accountability, states continue to move forward. New Jersey was an early leader in the statewide testing movement. In 1972, the state implemented the Educational Assessment Program (EAP). In 1975, the state constitution was amended to require the Legislature to establish a system of thorough and efficient education. The Legislature's response was the Public School Education Act to provide to all children in New Jersey, regardless of socioeconomic status or geographic location, the educational opportunity that would prepare them to function politically, economically and socially in a democratic society. An amendment to that act, signed in 1976, established uniform standards of minimum achievement in basic communication and computation skills. This amendment became the legal basis for

the use of a test as a graduation requirement in the state of New Jersey.

The outcome of the 1976 amendment was the Minimum Basic Skills Test (MBS). As of the 1981-82 school year, ninth-grade students were required to pass the MBS in reading and mathematics as a requirement for a high school diploma. The content of the test reflected the title.

In January 1983, New Jersey Commissioner of Education Saul Cooperman announced a new statewide testing system to replace the MBS. Also administered to ninth-grade students, the High School Proficiency Test (HSPT) became a graduation requirement as of the class of 1989 (New Jersey Department of Education, 1997b).

In 1988, the New Jersey Legislature passed a law that moved the High School Proficiency Test from ninth grade to eleventh grade (HSPT11). Following three years of due-notice testing, HSPT11 was first administered as a graduation requirement in October 1993 to all regular education eleventh-grade students (New Jersey Department of Education, 1989, 1990a, 1990b, 1997c). Since HSPT11 would be administered later in a student's high school career, it was important to identify at an earlier point those students at risk of not mastering the skills tested. Therefore, the 1988 HSPT law also established an eighth grade

Early Warning Test (EWT). The EWT was first administered in March 1991 (New Jersey Department of Education, 1989, 1997a).

In 1992, the state board of education mandated the establishment and administration of a statewide fourth-grade test in New Jersey Administrative Code (N.J.A.C.) 6.8-4.6(a)1. The fourth-grade test was intended to provide an "accurate measure of how elementary school students are progressing towards acquiring the knowledge and skills needed to graduate from high school and function politically, economically, and socially in a democratic society" (New Jersey Department of Education, 1999b, p. 5).

Although a fourth-grade test was mandated in 1992, implementation of the assessment instrument did not take place until 1997. Within that period, the New Jersey Department of Education engaged in a parallel task, the establishment of state academic standards. Adopted by the New Jersey state Board of Education in May 1996, the Core Curriculum Content Standards (CCCS) were intended to describe what all students should know and be able to do at the end of fourth grade, eighth grade, and upon completion of a New Jersey public school education. As would be expected in this world of accountability, the fourth-grade Elementary School Proficiency Assessment (ESPA) test specifications became aligned to the Core Curriculum Content

Standards. ESPA was developed in 1996 by National Computer Systems (NCS). The first administration, May 1997, was a practice test: ESPA was finally administered as an operational assessment in May 1999.

Paralleling the ESPA design and alignment to the standards, a new eighth-grade test was also developed. In March 1999, the Grade Eight Proficiency Test (GEPA) replaced the EWT. The third piece of the new state assessment program, the eleventh-grade High School Proficiency Assessment (HSPA) is currently being field-tested for a two-year period.

New Jersey Commissioner of Education David Hespe has signed on as a supporter of standards and testing. According to a December 1999 news release, the commissioner wrote:

> Today, every state but one has adopted higher academic standards and new assessments to prepare students for the challenges that await them in the new millennium. Every presidential candidate and just about every governor and legislature in the nation has endorsed standards-based assessments as the means to higher student achievement. New Jersey was in the forefront of this movement. In 1996, after considerable study and input from citizens and

educational practitioners, we established core curriculum content standards in seven subject areas and began working with education experts and New Jersey teachers on a new 4th and 8th grade test that would determine how well our children were performing...

At a seminar hosted by the Department of Education for school officials in our special needs districts, Dr. Simmons (Warren Simmons, executive director of the Annenberg Institute for School Reform at Brown University) said it is essential to link "rich standards" to "rich assessments." Critics who say the standards and assessment movement is robbing students of valuable classroom instruction by forcing teachers to "teach to the test" are missing the point. "If the assessments were rich enough, and authentic enough, and meaningful enough, then teachers could actually teach to the assessments" ... (another presenter) Dr. Ravitch told the convention of New Jersey school board members: "It's about the kind of society we are, the kind of people we want to be and about the

> role of education in raising all of our children to their highest potential."

While vigorously supporting the state testing program, Mr. Hespe's news release also acknowledged that there have always been doubters:

> ...in New Jersey and other states that are pursuing standards-based curriculum reform, special interest groups are questioning the need for higher standards and tougher assessments. They say the assignments are too demanding. They say the tests are too long. They say the results can not be right, so the tests must be flawed. What the critics are really saying is make the tests short and easy so everyone can feel good about the results. We realize that some parents and school officials may be upset by the results of the ESPA and GEPA and may seek to blame the messenger, which in this case is the new test. However, if we continued to administer the old test, which students were passing at better than a 90 percent rate, we would be doing our children a great disservice...

Among those questioning the state assessment program were two major professional organizations. On October 21, 1999, Daniel Money, cabinet member of the New Jersey Principals and Supervisors Association (NJPSA) presented testimony to the state board on behalf of the association. The testimony urged a delay in the implementation of the testing program until there was a careful review of reliability and validity. On an even stronger note, in January 2000, the directors of the New Jersey School Boards Association approved a seventeen-page series of recommendations on the state testing program covering everything from administration of the tests to their evaluation. Called a "scathing report" by The Star-Ledger (January 29, 2000), the recommendations were formally presented to the state Board of Education on February 16, 2000. Despite the concerns of educators, on April 5, 2000, the unanimous Legislature adopted the new Standards and Assessment Code - N.J.A.C. 6A (Mooney, 2000a).

Statement of the Problem

The issues of standards, assessments, and accountability are front and center across the nation. Of the three, the strongest debate is around assessment. Do state assessments represent the

modern Procrustean bed or are they viable and valid components of accountability reform?

Gillespie et al (1996) stated that the purpose of good assessment is to inform instruction and, at the same time, to provide students, parents, administrators, and the public with accurate and meaningful information about students' progress. For the information to be accurate and meaningful, the test must be reliable and valid. While reliability deals with the consistency of scores, validity addresses the very essence of the assessment. Simply stated, a test is valid if it measures what it is supposed to measure.

How important is validity? In December 1947, the Educational Testing Service (ETS) was chartered. On January 1, 1948, ETS opened for business. For the first two years of its existence, the new company confirmed the importance of psychometrics, confining itself to performing validity studies on those tests for which it held the copyright: SAT, MCAT, GMAT (Lemann, 1999).

Even those not intimately involved in psychometrics recognize the importance of validity. When considering the possibility of a national test, the executive committee of the American Association of School Administrators (AASA) endorsed the concept on the condition that the tests were valid (Lawton,

1997). Bob Chase, president of the National Education Association (NEA), mirrored this in a 1999 press release where he stated: When high stakes tests are the accountability measure for standards, it is important that they are valid.

According to Anastasi (1973, 1959), test validity revolves around what the test measures and how well it does so. In essence, does the test measure what it purports to measure? It is important to note that what a test measures is not necessarily reflected in the title or name. Rather, what is measured can only be determined by an inspection of the criterion. Mursell (1947) stated, "This may be considered a minimum requirement in proper test construction and evaluation" (p. 36).

The initial validity question regarding what the test measures leads to a second: can appropriate conclusions and inferences be drawn from the results? Cronbach (1971) stated that what is validated is not actually the test but rather the inferences derived from test scores. As noted by Moss and Schutz (1999), during the 1980s Cronbach "criticized the testing Standards for having a 'conformational bias' whereby 'validation consists not so much in questioning the proposed interpretation as in accumulating results consistent with it" (p. 686). Cizek (1998) agreed, calling validity "the degree to which the

conclusions yielded by the test are meaningful, accurate and useful" (p. 8).

The dilemma with validity is that the concept has transitioned from multiple, discreet formats such as face, content, etc., into the current unified concept of construct validity. Since the topic resides in the discipline of psychometrics, many educators have not kept pace with the changes. According to the Committee on Appropriate Test Use of the National Research Council Board on Testing and Assessment (1999), test validation is an empirical evaluation of test meaning and use. The committee further stated that the literature of psychometrics views the fundamental issue as construct validity based on the fact that the meaning of a test score is a construction. "Validity always refers to the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on test scores" (Messick, 1993, p. 1).

Construct validity is not a new term. Anastasi (1982) noted that construct validity was officially introduced into psychometrics in 1954 in the Technical Recommendations for Psychological Tests and Diagnostic Techniques, the first edition of test standards. Gronlund (1981) stated, "Broadly conceived, construct validity is an attempt to account for a difference in

test scores. Instead of asking, 'Does this test measure what the author claims it measures?' We are asking, 'Precisely what does this test measure?'" (p. 84). However, although validity has transitioned into a unified concept, Messick (1993) noted that construct validity can be divided into four aspects (content, substantive, structural, external). For the purposes of this study, the content and external aspects are the most relevant.

Content validity measures the degree to which the items on a test are a representative sample of the content domain assessed. In order for content validity to be accurate, the domain of tasks must be clearly defined and the subject matter content, including instructional objectives, should be identified. The latter describes the types of performance that pupils are expected to demonstrate (Gronlund, 1981).

It would be natural to assume that content validity is the most appropriate form for criterion-referenced state achievement tests. Through the referencing of standards and assessments in the same breath, that assumption is intensified. Rosenholtz (1991), Fullan and Steigelbauer (1991) stated that clear, intelligible standards are a pillar of higher achievement. Aligned with appropriate assessments, they can help us realize the dream of learning for all. This was echoed by Dr. Simmons

(1999) in his previously noted statement that it is essential to link rich standards to rich assessments.

The New Jersey Department of Education identified with this premise. An April 1, 1998, memo by former commissioner, Leo Klagholz, stated that the CCCS and the embedded progress indicators form the basis for curriculum, instruction and assessment in New Jersey schools. By defining what is to be learned, the standards are specific and measurable. The memo further stated that the New Jersey assessments are criterion-referenced and standardized to ensure reliability and validity.

According to Popham (1975), "because content validity involves someone's inspecting the items and deciding whether they are sufficiently consonant with the content or learner behaviors to be measured, there is obviously a heavy reliance on human judgment...Although the concept of content validity has been around for a good many years, there have been few exemplary applications of the approach" (p. 120). In reviewing content validity, Gronlund (1981) also noted that the procedure depends on logical analysis and judgment. He further stated that the adequacy of describing the content domain was a major concern.

Under content validity, it would be simple to state that the newer tests of the New Jersey assessment program (ESPA, GEPA, HSPA) are valid if they adequately measure the CCCS. For

content validity, the standards would be the content domain. The items on the test should then represent a measurable sample of that domain. If through logical analysis and human judgment, the items on the test strongly reflected (correlated to) the appropriate standard(s), the result would be a high content validity.

However, what does this mean? If the domain (criterion) is accepted as a good representative of the content area, high validity would indicate that the assessment is a good measure of the identified subject area. On the other hand, what happens if the domain is described by independent sources as a weak representation of the content area? Does a high content correlation between the assessment and the standards indicate anything more than the fact that the inferior domain and the assessment measure the same thing?

Analyses of standards are limited and often self-serving. The American Federation of Teachers (AFT), the Council for Basic Education (CBE), and the Thomas B. Fordham Foundation have each reviewed standards. Of these, the Fordham Foundation, based in Washington and headed by Chester E. Finn Jr., a former assistant US secretary of education under President Reagan, provided the best example of a consistent, structured and nationally recognized evaluation of state standards in five core academic

areas: English, history, geography, mathematics and science. According to The State of the State Standards reports (Finn, 1998, 2000), New Jersey scored an A only in science, with a C in mathematics, and ratings of F in English and history. This supported the 1996 American Federation of Teachers analysis which stated that New Jersey failed to meet its common core criterion and gave passing grades only to the state's mathematics and science standards (Mooney, 2000b).

Looking back to a July 1997 Fordham report on State English Standards, Dr. Sandra Stotsky, research associate at both the Harvard Graduate School of Education and the Boston University School of Education, wrote that the New Jersey Language Arts standards had "many limitations...Its standards for reading, literary study, and writing are weak...Many standards lack specificity and measurability, and they do not show much increase in complexity over educational levels." Her recommendation was that "The document needs to be completely rewritten...with specific and measurable standards " (p. 61).

The independent analyses of standards bring into question the feasibility of accepting content validity as an appropriate measure for ESPA, GEPA and HSPA. Content validity is only as good as the criterion against which the test is measured. It is apparent that questions exist as to the design and content of

the standards, particularly in language arts. Therefore, content validity could be high, but what would it tell us - that the tests accurately measure the test taker's knowledge of an imperfect set of standards?

For HSPT11, the current New Jersey graduation test, the question is different. HSPT11 was designed prior to the development of the CCCS. According to New Jersey Department of Education, the test was intended to be a rigorous assessment of essential skills in reading, mathematics and writing. The test purportedly measures the knowledge and skills expected at the completion of thirteen years of public education. Content validity for HSPT11 would compare the assessment to the published skills rather than to the CCCS.

Returning to the earlier stated definitions of validity, there are two aspects: 1. Does the test measure what it purports to measure? 2. Can appropriate conclusions and inferences be drawn from the results? If a state assessment purports to measure the CCCS, high content validity would indicate that it is an effective instrument. If the state assessment purports to measure reading, mathematics and writing, then the CCCS must first be proven to be true representatives of those content areas. In 1929, Carl Campbell Brigham, psychology professor at Princeton and assistant to Yerkes during WWI, wrote to Charles

Davenport, head of the Eugenics Record Office, "The more I work in this field, the more I am convinced that psychologists have sinned greatly in sliding easily from the name of the test to the functions or trait measured" (Lemann, p. 33).

There is another important point that must be addressed. High content validity can be used to effect the alignment of assessment, curriculum, and instruction throughout the state. However, content validity is not appropriate for impacting accountability. A 1993 professional article on validity by Messick, stated " ...in a fundamental sense so-called content validity does not qualify as validity at all, although such considerations of content relevance and representativeness clearly do and should influence the nature of score inferences supported by other evidence" (p. 17). Content validity provides judgmental evidence on the domain; it does not address test scores, the core of the validity concept.

To ensure appropriate conclusions and inferences for high-stakes assessments, the preferable format is external validity. Earlier known as empirical validity, Anastasi (1959, 1973) anointed it as the most important format.

Anastasi (1959) and Gronlund (1981) defined this aspect of validity as the extent to which test performance is related to another valued, independent and direct measure of that which the

test is designed to assess. Heubert and Hauser (1999) agreed, describing the external aspect of construct validity as the extent to which performance on a test is related to external variables. Convergent evidence should indicate that the test relates to other variables that theoretically measure the same thing. Quoting Loevinger, Messick (1993) wrote "The external component of construct validity refers to the extent to which the test's relationships with other tests...reflect the expected high, low and interactive relations implied in the theory of the construct being assessed' (p. 45). The external aspect of validity is an essential determinant when using test scores for accountability and high-stakes decisions.

In implementing the external approach, the most important ingredient is the quality of the criterion. According to Popham (1981), there have been many unfortunate examples where measurement people gathered correlation data relating test performance to a criterion variable that, under close scrutiny, was indefensible. To avoid this, a common format, supported by ETS and described above by Messick and Loevinger, is to correlate the test to another test already validated as representing the construct. High correlations would be expected between two tests in the same content area, for example two language arts tests or two mathematics tests. "The scores of any

particular test can be expected to correlate substantially with the scores of other tests that presumably measure the same thing...For any given test, we would predict higher correlations with like tests and lower correlations with unlike tests" (Gronlund, 1981, p.83). As Messick (1993) noted, "The external component of construct validity primarily concerns correlations with the test's total score and any subscores" (p. 45).

For this study, two additional validity terms must be addressed: concurrent and predictive. Concurrent validity provides the relationship between two measures obtained within a short period of time. The two sets of data are always on the same individual (Anastasi 1973; Gronlund, 1981). Correlating scores received by the same individual on two tests measuring the same content area, one of which is the state-developed test, is the preferable mode of determining the concurrent, external aspect of construct validity for the accountability functions of ESPA, GEPA and HSPT.

On the other hand, predictive validity is required when there is interest in predicting or determining the relationship between two measures over an extended period of time. Anastasi, one of the most noted psychometricians, stated, "If we want to use test scores to predict outcome in some future situation, such as an applicant's performance in college, we must use tests

with high predictive validity against the specific criterion" (1982, p. 30). As with concurrent validity, the two sets of data must always be on the same individual. Predictive validity answers the challenge of the EWT, and potentially the GEPA and the ESPA. The skills identified for the EWT were the benchmarks for those on HSPT11. In Department of Education meetings throughout New Jersey, it was stated that the test should be used to identify students who might have difficulty passing the high-stakes HSPT11. According to Gronlund (1981), "If the results are to be used to predict student success in some future activity, we should like them to provide as accurate an estimate of future success as possible" (p. 65). As with the EWT and HSPT11, the same relationship is expected between the GEPA and the, in development, HSPA as well as between the fourth-grade ESPA and the GEPA.

Minimal technical studies were conducted on the original HSPT (HSPT9). The conclusion was that, for a full study, the test would have to include secure items. When HSPT11 replaced HSPT9 as a graduation requirement, secure items were embedded. However, discussions with former New Jersey Director of Assessment, Gerald DeMauro, indicated that, while technical studies had been conducted on HSPT11, no true validation study was available. One study was conducted of the relationship

between EWT and HSPT11. The results were inconclusive and have not been released. With regard to the new assessments, two memos (1997, 1998) from Assistant Commissioner Ellen Schechter to Chief School Administrators noted that construct validity studies were conducted on the May 1997 ESPA. However, the studies were also unpublished. In addition, the state determined that the 1997 ESPA administration was sufficiently faulty that standard-setting was delayed until the 1999 administration.

Too often in education, practice follows theory without the interim step of appropriate research. Fenwick English, a national proponent of curriculum alignment and a professor of educational administration at Iowa State University (Ames), frequently refers to educators as lemmings, chasing every new idea without knowing whether it is valuable. This leads to a question for state assessments: Is there any proof that they are valid for high-stakes accountability?

Purpose of the Study

The purpose of this project was to conduct a validity study of the current New Jersey state testing program. It was noted that content validity may be a format for linking standards and instruction, dependent on the appropriateness of the content domain. However, it is not viable for accountability issues.

Therefore, content validity was eliminated as the acceptable format. Since the New Jersey state tests have accountability for students and districts, the external aspect of validity must be identified. Therefore, the purpose of this study was to determine the concurrent, external aspect of construct validity of three tests in the New Jersey state assessment program - Elementary School Proficiency Assessment (ESPA), Grade Eight Proficiency Assessment (GEPA) and High School Proficiency Test11 (HSPT11) - by correlating scores for the same students on the state tests and on grade and content appropriate, national, standardized tests. In addition, this study looked at the predictive validity of the EWT by correlating the scores for the same students on both the EWT and the HSPT11, the latter taken three years later in junior year of high school.

The first major research question was to determine if the current HSPT11 is a valid test with sufficient weight to warrant the use as a graduation requirement. It was recognized that the HSPT11 will be replaced by the High School Proficiency Assessment (HSPA), currently being field-tested. However, both tests claim to measure what should be learned at the completion of thirteen years of schooling; both were constructed, or are being constructed, by state department of education selected committees using the same development process; and the two are

coordinated by the same state department employee. Therefore, since no data are available for the HSPA, this study will research the validity of HSPT11. The information provided will be important in determining the future progress of the in-development HSPA.

The second major research question was to determine if the EWT did predict performance (predictive validity) on the HSPT11, laying a foundation for questioning if the GEPA can predict performance on the HSPA and, as an extension, if the ESPA can predict performance on the GEPA.

The third major research question was to determine if the GEPA, first administered in 1999, was valid when correlated to a recognized, independently validated, standardized test taken by the same students and measuring the same content areas.

The fourth major research question was to determine if the ESPA, first administered in 1997 with standards set in 1999, is valid when correlated to a recognized, independently validated, standardized test taken by the same students and measuring the same content areas.

The three recognized, national, standardized tests in this study are published under the auspices of ETS: SAT, Preliminary Scholastic Assessment Test (PSAT), Comprehensive Testing Program third edition (CTP III).

For the external aspect of validity, either concurrent or predictive, the indicated procedure is to correlate two sets of scores on the same individual and to report the degree of relationship between them by means of a correlation coefficient. "This enables validity to be presented in precise and universally understood terms" (Gronlund, 1981, p.73). A correlation coefficient of 1.00 indicates a perfect positive relationship while .00 indicates no relationship. The nearer the validity coefficient is to 1.00, the greater the accuracy in predicting one variable to another. According to lectures by Dr. Anastasi, for standardized assessment, the acceptable coefficient for validity is r=.7 or higher. A review of technical manuals for national tests supports this premise. In lieu of other research on the topic, to determine the specific research questions, r=.7 has been identified as the significant number.

Thirty specific research questions supported the four major research topics. The specific questions, tested as null hypotheses, allowed the division of the major research topic into correlations between aligned tests:

- Do the scores of the HSPT11 Reading test correlate .7 or higher with the scores of the same students on the PSAT Verbal test?

- Do the scores of the HSPT11 Reading test correlate .7 or higher with the scores of the same students on the SAT Verbal test?
- Do the scores of the HSPT11 Writing test correlate .7 or higher with the scores of the same students on the PSAT Verbal test?
- Do the scores of the HSPT11 Writing test correlate .7 or higher with the scores of the same students on the PSAT Writing test?
- Do the scores of the HSPT11 Essay test correlate .7 or higher with the scores of the same students on the PSAT Writing test?
- Do the scores of the HSPT11 Writing test correlate .7 or higher with the scores of the same students on the SAT Verbal test?
- Do the scores of the HSPT11 Mathematics test correlate .7 or higher with the scores of the same students on the PSAT Mathematics test?
- Do the scores of the HSPT11 Mathematics test correlate .7 or higher with the scores of the same students on the SAT Mathematics test?

- Do the scores of the EWT Reading test correlate .7 or higher with the scores of the same students on the HSPT11 Reading test?
- Do the scores of the EWT Writing test correlate .7 or higher with the scores of the same students on the HSPT11 Writing test?
- Do the scores of the EWT Writing Task (Essay) correlate .7 or higher with the scores of the same students on the HSPT11 Writing Task (Essay)?
- Do the scores of the EWT Mathematics test correlate .7 or higher with the scores of the same students on the HSPT11 Mathematics test?
- Do the scores of the GEPA Language Arts Literacy test correlate .7 or higher with the scores of the same students on the CTP III, Level E, Verbal Ability test?
- Do the scores of the GEPA Language Arts Literacy test correlate .7 or higher with the scores of the same students on the CTP III, Level E, Reading Comprehension test?
- Do the scores of the GEPA Language Arts Literacy test correlate .7 or higher with the scores of the same students on the CTP III, Level E, Writing Process test?

- Do the scores of the GEPA Reading component (Language Arts Literacy) correlate .7 or higher with the scores of the same students on the CTP III, Level E, Verbal Ability test?
- Do the scores of the GEPA Reading component (Language Arts Literacy) correlate .7 or higher with the scores of the same students on the CTP III, Level E, Reading Comprehension test?
- Do the scores of the GEPA Writing component (Language Arts Literacy) correlate .7 or higher with the scores of the same students on the CTP III, Level E, Verbal Ability test?
- Do the scores of the GEPA Writing component (Language Arts Literacy) correlate .7 or higher with the scores of the same students on the CTP III, Level E, Writing Process test?
- Do the scores of the GEPA Mathematics test correlate .7 or higher with the scores of the same students on the CTP III, Level E, Quantitative Ability test?
- Do the scores of the GEPA Mathematics test correlate .7 or higher with the scores of the same students on the CTP III, Level E, Mathematics test?
- Do the scores of the ESPA Language Arts Literacy test correlate .7 or higher with the scores of the same students on the CTP III, Level D, Verbal Ability test?

- Do the scores of the ESPA Language Arts Literacy test correlate .7 or higher with the scores of the same students on the CTP III, Level D, Reading Comprehension test?
- Do the scores of the ESPA Reading component (Language Arts Literacy) correlate .7 or higher with the scores of the same students on the CTP III, Level D, Verbal Ability test?
- Do the scores of the ESPA Reading component (Language Arts Literacy) correlate .7 or higher with the scores of the same students on the CTP III, Level D, Reading Comprehension test?
- Do the scores of the ESPA Writing component (Language Arts Literacy) correlate .7 or higher with the scores of the same students on the CTP III, level D, Verbal Ability test?
- Do the scores of the ESPA Writing component (Language Arts Literacy) "poem" correlate .7 or higher on the CTP III, Level D, Verbal Ability test?
- Do the scores of the ESPA Writing component (Language Arts Literacy) "picture" correlate .7 or higher on the CTP III, Level D, Verbal Ability test?
- Do the scores of the ESPA Mathematics test correlate .7 or higher with the scores of the same students on the CTP III, Level D, Quantitative Ability test?

- Do the scores of the ESPA Mathematics test correlate .7 or higher with the scores of the same students on the CTP III, Level D, Mathematics test?

Significance of the Study

Forty-nine states have developed content-area standards to describe what students should know in the respective areas. Forty-eight states have created assessments; of these twenty-four states have high-stakes tests. During 2000, forty states plan to issue school report cards based on test performance, twenty-one plan to issue overall ratings for schools, and eighteen expect to have the legal authority to close, take over, or create reforms in failing schools. Intensifying the stakes, in early February, Standard & Poor's, a division of the McGraw-Hill Companies, introduced School Evaluation Services. The service will allow states to link data about financial expenditures with academic results and compare performance across districts (Olson, 2000a). Both presidential candidates, Republican George W. Bush and Democrat Al Gore, have proposed tying federal education dollars to states' test scores (AP, 2000). As noted by Paul Barton, former president of ETS, testing has turned into a means of reform rather than the method of

determining whether the reforms have been effective (Sergiovanni, 2000).

That view of assessment is not likely to change. Public Agenda's Reality Check 2000 survey found that eighty-seven percent of employers, seventy-nine percent of professors, seventy-nine percent of parents and sixty percent of teachers favored high-stakes tests. Fifty-nine percent of employers and fifty-one percent of parents also agreed that high-stakes testing makes teachers and students more accountable.

As stated by Heubert and Hauser (1999), the use of tests for accountability has significant consequences for educators, schools, school districts and individual students by impacting school management, budgets and the quality of instruction. Yet, citing Stake (1998), the editors note that "it appears that many states have not taken adequate steps to validate their assessment instruments, and that proper studies would reveal important weaknesses" (p. 179).

In New Jersey, the assessment stakes are high, and getting higher. The current HSPT11 is a graduation test, as will be HSPA. N.J.A.C. 6:8-6.1 and 6.2 requires districts to provide individual comprehensive assessment for each student who scores below passing on one or more sections of HSPT11 and to develop

an Individual Student Improvement Plan (ISIP) which identifies the areas and strategies for improvement.

This becomes even stronger under the new code N.J.A.C. 6A:6-5.1. The new subchapter mandates that the awarding of a diploma be linked to developing the knowledge and skills contained in the Core Curriculum Content Standards as measured through the statewide system of assessment

In addition to the high-stakes involved for each high school student, accountability is also built-in for districts. According to former Commissioner Klazholtz, the state assessment system provides a foundation for certifying that local school districts and charter schools have aligned their programs to the Core Curriculum Content Standards and are providing a thorough and efficient education for all students. For the district to be certified, seventy-five percent of the eighth grade students in each school needed to pass the EWT in each subject area; for that same certification, eighty-five percent of the fourth grade students and eighty-five percent of the eighth grade students must pass each content area in the respective ESPA and GEPA. The same holds true for high school. Eighty-five percent of the appropriate students in each school must pass the HSPT11, and soon the HSPA, for the district to be certified. Loss of

certification has serious consequences, including potential state intervention.

District public relations are also impacted by the state assessment program. N.J.A.C. 6:39-1.4(a)3 requires each district, following a thirty day interpretation period, to make test results available to the public. That public accountability is strengthened through the state-issued New Jersey School Report Card that is provided to parents and made available to the community and media.

The impact of state assessments further extends to general curriculum and instruction. The state suggests that GEPA and ESPA scores should serve as "indicators for determining which local education programs may need revisions" (New Jersey Department of Education, 2000b, p. 1). Newspaper articles during the 1999-2000 school year note that many districts have completed major redesigns of curriculum based on the new standards and assessments, without any proof that either is effective or will lead to a better education. The revisions to curriculum and training of staff to deliver the revamped curriculum are cost factors that must be reflected in district budgets.

Adding extra pressure, on May 14, 1997, the New Jersey Supreme Court, in an adjunct to its order to state officials to

immediately increase funding for poor urban districts, stated that the New Jersey Core Curriculum Content Standards might eventually improve educational opportunity. "The content and performance standards define educational opportunity required by the Constitution. It is an effort that warrants judicial deference. We therefore conclude that the standards are facially adequate as a reasonable legislative definition of a constitutional thorough and efficient education" (Abbott et al, 1997; New Jersey Principals and Supervisors Association, 1998a). On May 21, 1998, the state supreme court reiterated that support for the standards. With the ESPA, GEPA and HSPA acknowledged by the state department of education as the assessments for the standards, the proof of a thorough and efficient education rests in these tests.

Across the nation, as in New Jersey, states claim to have conducted validity studies on their assessment programs. In New Jersey, early studies were conducted on HSPT9. These focused on the standard-setting procedure. For HSPT11, and the link between the EWT and HSPT11, the department of education stated that unpublished, non-public studies were conducted. On October 19, 1998, during a telephone discussion, Dr. Wendy Roberts, former manager of the EWT, stated that a technical manual for the test had been developed but was not available to the public. On that

same day, Dr. Eva Miller, also of the New Jersey Department of Education, stated that she and Dr. Gerald DeMauro, former director of the department of assessment, had conducted a linking study between the EWT and the HSPT. However, she noted that the data were inconsistent and will never be published. As for the HSPT, during a third October 1998 telephone discussion, Dr. Veronica Orsi stated that technical studies on the HSPT had been conducted annually, but none had been published for several years.

With regard to the new assessment program, two memos (1997, 1998) from Assistant Commissioner Ellen Schechter to Chief School Administrators noted that construct validity studies were conducted on the May 1997 ESPA. The memos stated that the studies revealed "substantial construct validity across the different types of test questions" (p. 4). Validity of test questions implies the content aspect of construct validity. As noted above, Messick (1993) stated, "...in a fundamental sense so-called content validity does not qualify as validity at all" (p. 17). He further noted, "Validity always refers to the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on test scores" (p. 1). Content validity provides judgmental evidence on the domain; it does not address test

scores. Returning to the Schechter memo, it is important to note that in spite of the purported, unpublished findings of "substantial construct validity," the 1997 ESPA raised sufficient concern to delay the standard-setting to 1999.

From Mursell through Anastasi to modern psychometricians such as Wiggins, validity is considered a minimum requirement in determining whether the conclusions and inferences are meaningful and accurate. However, also important is using the correct form. While this study could replicate a content validity study and be significant by making the results available to the public, questions raised by Messick as to the appropriateness of content validity as a true validity indicator coupled with those from the Thomas B. Fordham Foundation (Finn et al, 2000) about the rigor and appropriateness of the CCCS eliminated that as an acceptable format. In addition, as previously noted, content validity, when based on a good criterion, is a format for linking standards and instruction. However, it is not the preferable format for accountability issues. Since the New Jersey state tests have accountability for students and districts, the external aspect of validity must be identified.

This study will be the first to look at the external aspect of validity with regard to the New Jersey state assessment

program. It will also be the first published study to provide information on the following questions:

1. Is the HSPT11 a valid test to measure the skills and knowledge expected of students at the completion of thirteen years of public education?
2. Was the Early Warning Test an adequate predictor of performance on the HSPT? This has implications for the GEPA and ESPA. According to the New Jersey Department of Education, "the ESPA is intended to indicate the progress students are making...in mastering the knowledge and skills they will need to pass the GEPA. In turn, the GEPA is intended to indicate the progress students are making...in mastering the knowledge and skills they will need to pass the HSPA" (New Jersey Department of Education, 2000b, p. 1). The same statement was made for the EWT and the HSPT.
3. Is the recently developed GEPA a valid test of reading, writing and mathematics for eighth grade students?
4. Is the recently developed ESPA a valid test of reading, writing and mathematics for fourth grade students?

Definitions of Terms

Angoff: A method of scoring where knowledgeable judges rate each item on the test, estimating the proportion of marginal examinees who would correctly answer the question.

Assessment: The "gathering and synthesizing numerous sources of information for the purpose of describing or making decisions about a student...In education, the term assessment is increasingly used simply as a synonym for test" (Cizek, 1998, p. 11).

- Authentic Assessment: An assessment that replicates the challenges and standards of performance that face professionals such as writers, scientists, etc. (Wiggins, 1989). The emphasis is on the context in which the response is performed.
- Performance Assessment: A broad term covering assessments where students are required to demonstrate competencies or knowledge by creating an answer or product. This includes, but is not restricted to, constructed response items, essays, and portfolios (Feur & Fulton, 1993; Johns, 1992; VanHorn & Brown, 1983).

Correlation Coefficient: A number between -1 and 1 that describes the relationship between variables.

- Pearson Correlation Coefficient: Named for British scientist, Karl Pearson, this correlation coefficient describes the linear relationship between pairs of quantitative variables.

District Factor Group (DFG): A term used by the New Jersey Department of Education to indicate the socioeconomic status of citizens in each district and allow comparative reporting of test results from statewide testing programs. Seven variables were included in determining the DFG of each community/district: educational level, occupational level, density, urbanization, income, unemployment and poverty. The variables were combined with a statistical technique called principal component analysis, resulting in a single measure for each district. The DFG ranges from A (the thirty-five lowest socioeconomic districts) to J (the fifteen highest socioeconomic districts). All vocational districts were designated DFG "V". First developed in 1974 using demographic variables from the 1970 United States census, DFG was updated following both the 1980 census and the 1990 census.

Evaluation: The "ascribing value or worth to a score or performance. Saying that a student correctly answered 18 of 20 multiplication items is simply measuring the student's performance. Going beyond simple reporting to say the 18 of 20

correct should be judged to be proficient or awarding a grade of B+ represents evaluation of the student's performance" (Cizek, 1998, p. 11).

Item: Another word for a test question. If the test question is written in the multiple-choice format, the item has a *stem* (the part that introduces the question) and several *options*, usually labeled A, B, C, D, or similarly. If the test question asks the student to engage in a performance or demonstrate a skill such as writing an essay, the item is called a *prompt*—the written or other material introducing the task" (Cizek, 1998, p. 7).

Objectivity: A test is objective to the degree that a number of different individuals administering it to the same group and scoring it will arrive at results that agree. To achieve this, test developers often rely on a single correct answer per item.

Reliability: A test is reliable when there is a consistency of scores obtained by an individual who takes the same test on different occasions or who takes two tests with different but equivalent items. A test can be reliable without being valid.

Regression Analysis: Through data, relationships are identified among variables. The relationships can be used to make predictions.

- Linear regression measures the constant rate of increase of one variable with respect to another (Glasserman, 1999; Levine et al, 1999).

Scale Score: Scaling is the process of connecting numbers with performance on a specific test to indicate increasing levels of achievement. A score scale allows equating between alternate and/or subsequent forms of an assessment (Petersen, Kolen, Hoover, 1989).

School Report Card: Published annually for each public school in the state, the New Jersey School Report Card presents a statistical picture of the school through five main sections: school narrative; school-level demographic and organizational data; school-level student achievement results for relevant state tests as well as results, where appropriate, of the SAT and Advanced Placement examinations; information, where appropriate, about high school graduates; district-level personnel and fiscal information.

Standardized: A standardized test is one in which each condition impacting performance i.e., procedure, apparatus and scoring, are specified so that precisely the same test can be given at different times and places. For scoring, this theoretically eliminates interpretive errors.

Test: An objective and standardized measure of a sample of behavior. The term is used regardless of the format of the assessment (multiple-choice, true/false, performance, oral examination, essay, etc.).

- Ability test: A test on which the test taker is encouraged to earn the best score possible for that individual.
- Achievement test: One that measures knowledge or material learned in a content area.
- Aptitude test: A test used to predict success in a given field or subject area.
- Criterion-referenced Test (CRT): CRTs measure whether a student knows or can do specific things. Criteria for success are based on judgment. The score, usually pass or fail, indicates that a student has/has not met the criterion. It should be noted that the score does not indicate whether the criteria represent noteworthy expectations, whether the content is challenging, nor the whether the outcomes measured are desirable (Cizek, 1998).
- High-stakes test: A test used to make important decisions about students, teachers, and schools.
- Norm-referenced Test (NRT): Constructed to cover content that is generally considered appropriate for a grade level and/or content area, NRTs compare the performance of each

student to that of a normative group. The normative group should be a stratified, random sample covering all possible variables. Development of the normative group is usually based on the most recent census data.

- Proficiency Test: A test that measures the ability to perform a reportedly significant task.
- Standards-referenced Test (SRT): According to Cizek (1998), an SRT is similar to a CRT in that both attempt to describe the knowledge, skill or abilities that students possess. Where CRTs express standards in terms of quantity and category, SRTs link students' scores to concrete statements about what performance at the various levels means. Typically, SRTs are constructed to match content standards, academic statements of what students should know and be able to do in specific subjects or across several subjects. Performance could be described in levels such as Beginning, Proficient and Expert, each of which is linked to specified content performance. An example of an SRT is the National Assessment of Educational Progress (NAEP) which reports students' performance as Basic, Proficient and Advanced.

Validity: Validation is scientific inquiry into the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions

based on test scores (Messick, 1993, pp. 1,2). A test is valid if it measures what it is supposed to measure and the conclusions are accurate and useful. While a test may be reliable and not valid, the reverse is not true. A valid test is a reliable test. According to Bailey (1987) if a measure is valid it will be accurate every time and therefore must also be reliable. Since 1985, the American Educational Research Association has viewed validity as a unified concept. Although unified, current researchers include the following aspects of construct validity: content, substantive, external, and structural. Content and external are the most appropriate forms for this study; only these two aspects will be addressed:

- Content: The content aspect of construct validity refers to the extent to which test content is an adequate representation of the domain to be measured.
- External: This measures the extent to which performance on a test is related to external variables. For example, two tests of mathematics computation at the same grade level should show a strong correlation. In situations emphasizing accountability, whether selection, placement, graduation, certification and/or program evaluation, this aspect of validity is important.

Limitations of the Study

The study was conducted in one school district. The district was selected for the following reasons:

- The district was classified by the New Jersey Department of Education as a member of District Factor Group (DFG) I, the most numerous DFG with 105 districts.
- The district was recognized by Governor Christine Todd Whitman in her 1996 State of the State address as providing an outstanding education at less than the average per pupil cost, thereby validating the curriculum and achievement levels already in place.
- The high school has repeatedly been identified by New Jersey Monthly as one of the top ten in the state, indicating that the current curriculum covers the skills and knowledge expected at the completion of thirteen years of public education.
- One hundred and sixty members of the Class of 1999 took the PSAT, SAT and HSPT11 while students in the district; one hundred seventy-three members of the Class of 1999 remained in the district between the EWT and the HSPT11. This provided a broad data base.
- While an April 1, 1998 memo from then education Commissioner Leo Klagholz to the state board of education

noted that the state no longer required districts to administer assessments other that those in the state testing program, the district board of education opted to require a norm-referenced, standardized test for fourth grade students. This provided the opportunity to correlate the ESPA to a recognized standardized test.

The study contains the following limitations:

- The study included one district. Therefore, conclusions may not apply to other districts.
- The population included only regular education students.
- The data reflected the entire regular education population. There was no disaggregation of minority data. It should be noted that, on average, approximately twenty-two percent of the regular education student population are listed on school registration forms as members of a minority group.
- For the predictive validity study, the population included only students who were in the district for both the administrations of the EWT and that of HSPT11. Therefore, mobility was not a variable in this study. It should be noted that the district mobility index ranged around four percent. This is consistent with the average mobility for DFG "I" districts.

- The study only addressed the New Jersey state assessments. Data and conclusions may not generalize to assessment programs in other states.

Organization of the Study

The study is organized into five chapters. Chapter I includes the introduction, statement of the problem, purpose of the study, significance of the study, definition of terms, limitations of the study, and the organization of the study. Chapter II provides a review of the literature related to the study including a history of testing, the SAT, achievement testing and the Educational Testing Service; a review of selected state assessments - New York, New Jersey, California, Florida; the impact of standards; a discussion of problems, accountability, and consequences connected to assessment; an overview of the Voluntary National Tests, the National Assessment of Educational Progress, and the Program for International Student Assessment; a discussion of the psychometric considerations of assessment with an emphasis on validity; and a summary. Chapter III describes the data population, assessment instruments including those of the 1999 New Jersey state testing program (Eleventh Grade High School Proficiency Test - HSPT11, Early Warning Test, Grade Eight

Proficiency Assessment and Elementary School Proficiency Assessment) and selected tests from the Educational Testing Service (SAT, Preliminary Scholastic Aptitude Test - PSAT/NMSQT, and Comprehensive Testing Program third edition), data collection and analysis procedures and concludes with the study design for each of the four major research questions. Chapter IV presents an overview of the validity study, a statistical analysis of the quantitative data related to each of the four major research questions, and a summary of the findings. Chapter V includes the summary, conclusions and recommendations of the study.

Chapter II

REVIEW of the LITERATURE

Historical Perspective

In the testing movement, everything old is new again. The first assessments, biannual pupil exams, were instituted in 1122 B.C. by the Chow Dynasty. In China, testing continued through the Han, Tang and Sung dynasties. Under the latter, sophisticated methods were introduced. In 922 A.D., the Sung Dynasty improved scoring objectivity by removing candidate's names and assigning ID numbers. Between 960-1299 A.D., the Sung Dynasty introduced the practice of a neutral individual recopying exams so that handwriting did not influence the scores. During that same period, two independent readers for written exams became the norm. Unfortunately, in 1905, the same year that Alfred Binet and Theophile Simon introduced the first individually administered intelligence assessment and became known as the fathers of the testing movement, the Manchu Dynasty abandoned China's historic and precedent-setting exam system (Nitko, 1983).

Assessment history in the United States began in 1845 when Boston became the first district to print required short-answer tests (Hoff, 1999b). Although established in 1784, the New York Board of Regents did not enter the testing field until 1865 when

it introduced exams for elementary students. This was followed in 1878 with exams for secondary students (Nitko, 1983).

The concept of accountability through the use of annual exams to evaluate school programs began in 1562 with the Merchant Taylor's School in London. That process moved to the United States in 1856 when the Chicago public schools introduced written exams for elementary promotion and admission to high school. In 1897, Joseph Mayer Rice introduced the use of tests to evaluate schools. Rice published the results in The Forum.

Although Rice was the first in the United States to use assessment to evaluate schools, he did not meet the accountability challenge of Samuel King. In 1870, Samuel King, first superintendent of Portland, Oregon, published exam scores of pupils in newspapers. In 1877, parent and teacher outrage forced his resignation (Nitko, 1983).

From Intelligence Tests to the SAT and the ETS

Many psychologists look to Alfred Binet and Theophile Simon as the true fathers of the testing movement. In 1905, the French duo facilitated the first individually administered intelligence test. The test was intended to identify slow learners so that they could receive assistance. In 1916, Lewis M. Terman of Stanford University published the United States version, the

Simon-Binet Scale of Intelligence. While Binet had used the term mental age (MA), Terman introduced the term intelligence quotient (IQ). Along with Edward Thorndike of Columbia University, he became an advocate of widespread IQ testing to assess, sort, and then teach students in accordance with their abilities (Lemann, 1999).

In 1917, intelligence testing became accepted when Robert Yerkes, a Harvard professor, convinced the Army to allow him to administer IQ tests to almost two million recruits as a way of choosing officer candidates. Through the Army Alpha and Beta, the IQ movement was able to build a record of statistical evidence. It should be noted that the Army Alpha was developed in Vineland, NJ. In a parallel move, also in 1917, Arthur Sinton Otis made available the first, commercial, group intelligence test (Nitko, 1983; Lemann, 1999). Shortly after this, Lewis Terman devised a "National Intelligence Test" for elementary school students which became so popular that more than half a million students took the test annually during the 1920s (Lemann, 1999).

The Army Alpha did not die with World War I (WWI). Carl Campbell Brigham, a psychology professor at Princeton and assistant to Yerkes during WWI, continued to administer a version of the test to Princeton freshman and applicants to New

York City's Cooper Union, an all-scholarship technical college. Throughout the 1920s, similar, related studies were taking place around the country. E. L. Thorndike at Columbia University constructed an intelligence test for students at Columbia and the University of Pennsylvania. Yale also administered intelligence tests for students (Lemann, 1999).

Brigham's work with the Army Alpha included upgrading the questions. By 1926, the test had become the SAT. Originally known as the Scholastic Aptitude Test, briefly called the Student Assessment Test, the SAT is now known officially by its initials. At the outset, the SAT score was a single number. Brigham was persuaded by an assistant to divide the SAT score into two parts, one for verbal and one for mathematical ability.

Due to Brigham's contacts at the College Board, he was able to arrange for formal administration of the SAT soon after its development. Charles W. Eliot founded the College Entrance Examination Board (CEEB) in 1899 as an association for a few dozen private schools and colleges. In 1901, the first CEEB entrance examination was implemented. The College Boards was a uniform admissions test desired by boarding schools as one that all colleges would accept. On their side, the colleges wanted to impose some curricular order on the schools so that their

students would arrive with a consistent background (Lemann, 1999).

According to Lemann's history of the SAT (1999), Brigham viewed the test as an exam that Ivy League schools could use to award scholarships to students who did not come from elite New England families. The official date of the first SAT was June 23, 1926. Under the auspices of the College Board, 8,040 high school students took the test and had their scores reported to colleges. The Army also allowed the SAT to be used to test applicants to West Point; in 1930, the Navy did the same with Annapolis. Applicants to Yale and Princeton took both the College Boards and the SAT. Yale Law School adapted the SAT into a test for all applicants. In true psychometric fashion, the SAT was being used in each case not to decide who was admitted but to build up a validity record by correlating scores with the applicant's freshman grades.

As pictured by Lemann (1999), on the Sunday in December 1941, when the Japanese bombed Pearl Harbor, a group of College Board officials happened to be having lunch together in Princeton. During lunch they discussed abolishing the original College Boards essay examinations and using the SAT for all applicants, not just scholarship students. Within two weeks, the

essay examinations had been suspended for the duration of the war. They were never resumed.

In December 1947, the Educational Testing Service (ETS) was chartered with Henry Chauncey as president and Harvard president James Bryant Conant as chairman of the board. On January 1, 1948, ETS opened for business.

ETS combined the testing activities of the College Board, the American Council on Education, and the Carnegie Corporation of New York under one umbrella. Although ETS was established as a private organization performing a quasi-public function, its mission was to research and promote tests. The corporation had the advantage of owning the copyright to all the most prominent tests in higher education. An interesting point is that Carl Campbell Brigham, inventor of the SAT, had been so adamantly opposed to the creation of ETS that it could not have been created until he died (Lemann, 1999). Now, according to Donald M. Stewart, a former College Board president and currently a senior program officer for the Carnegie Corporation, "The relationship between the College Board and ETS is so much centered on the SAT, with the two organizations having in effect joint ownership of the test, that it would be impossible to break that relationship apart" (Hoff, 1999f, p. 33).

During World War II (WWII), the College Board became the testing consultant to the Navy. The Navy established a program called V-12. Young men with potential to perform advance technical jobs were recruited and sent to college for training. Henry Chauncey viewed the V-12 as a chance to demonstrate that a multiple-choice test could be given under secure conditions at many sites at once with quick reporting of the scores to a central authority (Lemann, 1999). This became a major initiative that fostered the growth of testing in general and the SAT in particular. Once it could be given on a mass scale for little cost, the test became a standard hurdle for college admissions.

Computerized score sheets further allowed the SAT to grow to half a million test-takers by the 1959-60 school year, then to 1 million just five years later. Today, the SAT is given almost 3 million times a year to more than a million students (Hoff, 1999f, p. 29). According to Schwartz (1999), the SAT has become the "single most important test for American high-school students - an academic and psychic rite of passage that strongly influences future educational options..." (p. 30).

In 1954, John Stalnaker became involved with the National Merit Scholarship Corporation. The concept was to give a national test to high-school students and, based on the score, a few students would be awarded full four-year college

scholarships. Stalnaker contacted ETS to devise the National Merit test. Eventually the test ended up in the hands of ETS where it remains today. Combined, the SAT and the Preliminary Scholastic Assessment Test (PSAT)/National Merit Qualifying Test (NMSQT) are responsible for 5 million tests a year.

As ETS blossomed, so did the critics. In 1957, John Gardner, head of Carnegie and an ETS board member, authored an article for Harper's that warned of the tyranny of testers (Lemann, 1999). His fellow board members did not appreciate his comments. However, he was not alone with his questions.

One of the more vocal ETS critics was Banesh Hoffman, a professor of mathematics at Queens College in New York, and a physicist and amateur grammarian. As recounted by Lemann (1999), in 1956, Mr. Hoffman wrote a critique of the SAT, based primarily on grammatical quibbles. He followed this with a 1962 book, Tyranny of Testing, and became a familiar figure on public television. Hoffman thought that the SAT discriminated against those with unusually high IQs. While this did not catch on, his disdain for the ETS psychometricians did.

Ralph Nader, a graduate of Princeton and Harvard Law School, famous in the 1960s for his crusade against automobile companies and federal agencies for their inattention to public safety, also joined the naysayers. In the early 1970s, after

reading the Tyranny of Testing, Nader started working criticisms of ETS into his speeches on college campuses. In the spring of 1974, he spoke at a community college in New Jersey. Attending that lecture, Allan Nairn, a high-school senior and future Princeton student, approached Nader to propose that they become a team and mount an investigation of ETS. Nader agreed. Nairn's work, coupled with Nader's influence, led to a Federal Trade Commission investigation of the test-preparation industry.

By the late 1970s, bills to regulate ETS were pending in thirty-seven state legislatures. The bill that hurt the most was the New York truth-in-testing bill signed into law by Governor Hugh Carey on July 13, 1979. This law forced ETS to release past tests to the public. Although only a New York law, ETS decided to treat it as federal legislation because it would be too difficult to develop one non-secure test for New York and one permanently secure test for the rest of the nation. That bill became the hallmark for the national publication of previously administered SATs (Lemann, 1999).

Standardized Achievement Tests

Paralleling the transition of intelligence tests into the SAT was the growth and adaptation of achievement tests to the demands and criticisms of society. During the early 1900s, an

interest in standardized achievement tests became the rage. In 1908, C. W. Stone developed the prototype. In 1911, P. H. Hanus facilitated the concept of school surveys. This spurred an interest in standardized achievement test results. The following year, 1912, tests were given further credence when D. Starch and E. C. Elliot showed that teachers' grades were unreliable. In 1914, the World Book Company made available the first commercial standardized achievement test: Courtis Standardized Research Test in Arithmetic (Nitko, 1983).

In 1923, Lewis Terman developed the Stanford Achievement Test. Still in existence, and occasionally known as the SAT, an acronym confusion, the original intent was to measure student achievement in grades two through eight. As reported by Hoff (1999b), Terman developed a national sample of 350,000 students, the largest to that point. This allowed a comparison of student achievement.

In 1927, William S. Learned and his partner, Ben D. Wood, head of the department of statistics at Columbia University Teachers College, developed standardized achievement tests that they administered through the Pennsylvania Study for the Carnegie Foundation for the Advancement of Teaching. Dr. Wood constructed similar objective exams for high school students under the auspices of the New York State Regents.

Through their testing, Wood and Learned discovered that there was no uniform curriculum nor grade-level, content-area expectations. In Pennsylvania high schools and colleges, all that a student needed to earn a diploma was to prove that he or she had attended class. No one in Pennsylvania, New York, or anywhere in the United States, had endeavored to find out what, if anything, students were learning. As a result, Learned and Wood proposed to publish a body of material that all students in high school and college should be required to master. After testing students on the mastered curriculum, those who did not succeed would be removed from the student population. The end product would be to reward students for mastery of a body of knowledge. Although the word meritocracy would not be invented until 1958 by Michael Young, head of research for the British Labour Party, the germ of the concept was present - "a social order that society rewards those who deserve and have earned advancement, rather than distributing reward by circumstances of birth" (Lemann, 1999, p. 342).

In 1927, while making a name as a leading promoter of standardized tests, Dr. Wood also founded the Educational Records Bureau (ERB) as a not-for-profit membership organization to develop educational assessment instruments. "At the time, Dr. Wood's concern was that the instruments available did not

provide adequate measures for high ability, high performing students in the independent and suburban schools of the country" (ETS, 1995, p. 1). A descendent of these tests, the Comprehensive Testing Program third edition, is one of the assessments used in this research study.

During this same period, E. F. Lindquist, a psychology professor at the University of Iowa, concentrated on elementary achievement tests. In 1929, he developed the Iowa Every Pupil Examination, followed in 1943 by the Iowa Tests of Educational Development and in 1945 by the Test of General Educational Development (GED). In 1936, in an interesting move, Lindquist condemned the use of tests to select gifted students or to impose uniform standards (Lemann, 1999). Lindquist would later (1959) create American College Testing (ACT), the strongest rival to ETS and the SAT. ETS's base was the East; ACT's the Midwest. In the early days, ETS tests primarily measured aptitude; ACT measured achievement.

Linquist further spurred the growth of achievement tests with the 1955 development of a high-speed, high-volume digital test-scoring machine. This was not the first scoring machine. Ben Wood had worked with Thomas Watson, the founder of IBM, to develop a machine that could score thousands or even millions of tests in the mass administrations that they envisioned. In 1936,

that IBM machine, based on the design of a former science teacher, Reynold B. Johnson, was used to score tests for the New York Regents. In 1968, Linquist further refined his 1955 concept by designing a machine for scoring test booklets (Lemann, 1999).

Achievement testing in the 1970s tended to follow the minimum basic skills or minimum-competency principle. This occurred at both the state and the national levels as Title 1 of the Elementary and Secondary Education Act (ESEA), first passed in 1965, stressed evaluation of basic skills. ESEA was, and is, the main federal law impacting K-12 schools. As reported by Hoff (1999b), between 1973 and 1983, "the number of states with minimum-competency tests grew from two to 34" (p. 26).

In 1983, A Nation at Risk was published. The federal report denounced the nation's schools and called for strengthening of the core curriculum and raising expectations using measurable standards (Berman, 1999). According to Hoff (1999b), "In the late 1980s and early 1990s, some states experimented with 'authentic assessments' based on portfolios of student work and questions that required them to write essays on exams" (p. 26). Conservatives and traditionalists fought this concept. In time, California dropped its program while Vermont and Kentucky added standardized tests. This left Maryland as the only state to rely solely on a performance assessment system.

As with the SAT, criticism arose around achievement testing. As early as 1880-1890, elementary school promotion exams came under heavy fire from teachers and other educators (Nitko, 1983). Moving to the 1950s and 1960s, professional educators frequently wrote articles advocating the avoidance of testing excesses. According to Lemann (1999), "Testing, in the 1950s, was still a new, raw business, full of bare-knuckled players and shoddy practices. All over the country, former school superintendents were setting themselves up as test publishers to make a little money. They sold tests with no validity or reliability studies to assure the quality. Test security was negligible" (pp. 96-97).

Criticism continued into recent decades. In October 1985, Boston College's George Madeus was one of over 900 educators, psychologists, and test constructors who attended an invitational conference in New York City organized by the Educational Testing Service (ETS). Following the conference, Madeus expressed concern that performance on tests was becoming what was most valued in education, that tests would come to dominate curriculum, and that tests could manipulate learners (Smith, 1986). Madeus's words echoed those of Carl Campbell Brigham, developer of the SAT. On January 3, 1938, Brigham wrote the following to James Bryant Conant, president of Harvard, "If

the unhappy day ever comes when teachers point their students toward these newer examinations, and the present weak and restricted procedures get a grip on education, then we may look for the inevitable distortion of education in terms of tests. And that means that mathematics will continue to be completely departmentalized and broken into disintegrated bits, that the sciences will become highly verbalized and that computation, manipulation and thinking in terms other than verbal will be minimized, that language will be taught for linguistic skills only without reference to literary values, that English will be taught for reading alone, and that practice and drill in the writing of English will disappear (Lemann, 1999, pp. 40-41).

The current emphasis on accountability has engendered similar criticism. In 1999, resistance was noted during a poll of 1,075 K-12 teachers (with oversampling in Florida, Massachusetts, New York and Texas) conducted by the Washington-based Peter D. Hart Research Associates for the Albert Shanker Institute, a non-profit organization formed in 1998. Seventy-three percent of the respondents agreed with the drive to raise academic standards. However, many expressed concern about the accompanying testing requirements. More than half of the teachers said that focusing on tests had narrowed the curriculum and omitted important areas. More than sixty percent said too

much time was spent on test preparation and thirty-nine percent said testing was too frequent (Bradley, Hoff & Manzo, 1999).

In an unusual and dramatic move, the Board of Directors of the International Reading Association (IRA) adopted a May 1999 position statement opposed to high-stakes testing. Published in the November 1999 The Reading Teacher, the statement noted the association's concerns with the reliance of both policy makers and educators on high-stakes testing.

> ...testing has become a means of controlling instruction as opposed to a way of gathering information to help students become better readers...
>
> There are several possible problematic outcomes of high-stakes testing. These include making bad decisions, narrowing the curriculum, focusing exclusively on certain segments of students, losing instructional time, and moving decision making to central authorities and away from local personnel...high-stakes tests have a tendency to narrow the curriculum and inflate the importance of the test.
>
> Less positively, politicians, bureaucrats, and test publishers have discovered that they

> can influence classroom instruction through the use of high-stakes tests. Tests allow these outside parties to take control away from local educational authorities without assuming the responsibilities of educating the students (pp. 257, 259-260).

The Voluntary National Tests

With assessment and accountability an important focus across the United States, it is no surprise that testing almost went national. In his 1997 State of the Union address, President Clinton proposed a national testing program. According to his vision, fourth and eighth grade students throughout the country would take the exams in the spring of 1999 (Olson, 1997). At the time of his address, President Clinton and his advisors assumed that the National Assessment of Educational Progress (NAEP) could be used for the national test with some minor modifications. The NAEP tests began in 1970, first collecting long-term data in science. In the early 1970s, mathematics and reading tests were added. The large, random, stratified sample, consistent demographic makeup and expansive range of questions made NAEP an attractive choice for the national assessment.

However, the concept of incorporating NAEP as the base for the national tests turned out to be an error in judgment. Experts claimed that the two, forty-five minute sections of the NAEP would not generate enough questions to pinpoint an individual student's achievement level. Other questions were also raised. These ranged from large issues, such as how to determine an individual score, how to accommodate Limited English Proficient (LEP) students, and how to integrate the 1997 Amendments to the Individuals with Disabilities Education Act (IDEA 1997) which required all states to include students with disabilities in statewide and districtwide educational assessments [Section 612(a)(17)(A)], to smaller management issues, such as how to purchase and appropriately distribute the calculators needed for the eighth grade test.

In March (Hoff, 1997b), a group representing the states' top education officials endorsed the outline of the national testing proposals, although several members expressed doubts that the final product would be what they need. The Council of Chief State School Officers adopted a one-page resolution supporting the plan to open new ways for students to strive toward world-class performance. In a full day of debate over the proposed resolution at the organization's annual legislative conference, members identified three areas still to be

addressed: avoiding duplication with existing state tests; ensuring the reliability of the tests during the fast-track development scheduled envisioned by the administration; and keeping costs down.

In June 1997, Michael Cohen, chief adviser to President Clinton on the national testing initiative, stated that only local officials would see individual scores. While the tests would be based on the National Assessment of Educational Progress (NAEP), private testing companies would license the right to sell the national tests with a battery of their own assessment. The publishers would score the tests and forward the scores to the local officials, the same procedure currently followed for district testing programs. The US Department of Education would not be involved in scoring, calculating or reporting test data. Mr. Cohen also stated that the federal government intended to pay testing costs for the first year. A March 19, 1997 letter from Marshall S. Smith, then acting deputy secretary of education, to congressional Republicans, estimated annual cost of the national tests to be 10-12 million dollars. Money for the creation of the tests was expected to come from an existing fund for the improvement of education under the aegis of the office of educational research and improvement (Lawton, 1997a). In future years, a charge of six to eight dollars per

student would be levied. Ramon C, Cortines, the acting assistant secretary for educational research and improvement at the Department of Education and former New York City Superintendent of Schools, was appointed to organize a team of testing experts to write the tests (Hoff, 1997c).

On September 11, 1997, the Senate voted 88-12 to give a nonpartisan National Assessment Governing Board (NAGB) full authority over the proposed tests. The following week, on September 16, 1997, in an appropriation amendment by Rep. Bill Goodling (R-PA), the House voted 295-195 to deny the Education Department the authority to spend any money on the tests. Senator John Ashcroft, R-MO, then tried to rally support to stop the testing plan on any terms. By the middle of October 1997, thirty-five GOP senators had banned together to block any further testing initiative (Hoff, 1997d).

The loss of funding hurt. On September 25, 1997, Secretary of Education Richard W. Riley announced that he was suspending test development in concert with the White House. In addition to the lack of support in Congress, the test was further weakened by a conflict over the use of a calculator in the math test and by the decision to administer the reading test only in English. The former was considered "emblematic of a schism nationwide about whether basic skills or higher order thinking should hold

sway in math classes" (Lawton, 1997c, p. 1). The latter angered many urban officials causing them to drop plans to administer the test. These districts included Houston, El Paso and Los Angeles, all of which have sizable populations with limited proficiency in English.

By the end of October 1997, Representative Bill Goodling (R-PA), chairman of the House Education Committee, still held firm to his opposition to proposed voluntary national assessments for fourth graders in reading and eighth graders in mathematics. He recruited 295 members to support his amendment to bar the Department of Education from spending money to develop the tests. Conservatives and liberals aligned in a House vote to block the plan, only to be overruled later in a compromise with the administration that allowed a nonpartisan board to continue with test development while researchers and Congress continued to debate the merit of national testing. According to Rep. Mark Souder, R-IN, "for our (conservative) base, this issue is World War III" (Hoff, 1997d, p. 24; 1997e).

By August 31, 1998, at an educational round table, President Clinton was silent on the issue of testing as a continued priority (Hoff, 1998a). That silence rang bells. On October 7, 1998, Rep. Goodling delivered the obituary, "It's

dead, dead, dead. There will be no national testing" (Hoff, 1998c, p. 22).

National Assessment of Educational Progress

While Clinton's national testing plan did not make it through the political wars, the concept of national testing may still find victory. Supporters carry on with the fight believing that, as the Council of Chief State School Officers stated in their resolution for the national test, assessment can "open new ways for students to strive toward world-class performance" (Hoff 1997b, p. 19).

Some supporters of national testing believe that the solution is NAEP. In 1963, U.S. Commissioner of Education Francis Keppel formed a panel of experts to explore the creation of a national test. The committee, led by Ralph W. Tyler, recommended a "regular sampling of students in basic subjects... Because local school administrators objected to reporting a breakdown of scores by state, Tyler's panel proposed that the scores be reported by four regions...(this) meant NAEP results would be useless for evaluating the effectiveness of schools" (Hoff, 1999b, p. 26).

The NAEP tests began in 1970, first collecting long-term data in science. In the early 1970s, mathematics and reading

number of students who attempted the item" (Cizek, 1998, p. 8). This allows for the development of a final version of the test that includes a range of items to challenge each level of learner. For national assessments, items are field-tested nationally; for state assessments, items are field tested across the state.

During the 1980s, CTB/McGraw-Hill capitalized on the fact that test development followed the same process regardless of the final intent for the product. With that in mind, the company marketed its tests, the California Achievement Test (CAT) and the Comprehensive Test of Basic Skills (CTBS), for use as both norm-referenced assessments and as criterion-referenced formats. For the latter, the company stressed the term "mastery." The dual use, coupled with the concept of mastery, skyrocketed CTB/McGraw-Hill into the premiere position among test publishers of that decade. A recent article by E. D. Hirsch Jr. (2000) supported this premise. Dr. Hirsch suggested that norm-referenced, standardized reading tests can also be used as competency-based tests when using the scale (absolute) scores. He further states that "All of the well-established reading tests are valid, reliable, and highly correlated with one another" (p. 40).

A lead teacher in the district of the data population serves on both a New Jersey state assessment committee (ESPA) and a national test development committee (CTP IV). She has repeatedly remarked on the similarity of the procedures. This echoed Popham. As quoted by Hoff (1999b), Popham stated, "More often than not, the (state tests) look like warmed-over versions of standardized (norm-referenced) tests...The mentality (test publishers) bring when they create a test is what they know" (p. 27). Popham and Hoff further point out that the Texas test was developed by Harcourt Educational Measurement and that modified versions of Harcourt's Stanford-9, a successor to the original Terman test, are used in several states including California. At the same time, CTB/McGraw-Hill is the contractor for the current Kentucky assessment system.

Whether the test is norm-referenced, criterion-referenced, or standards-referenced, a second consideration focuses on the choice between traditional multiple-choice questions and alternative response formats. Many accept at face value the fact that performance assessments are "better than traditional multiple-choice tests, better matched to new theories of curriculum and learning, and are more suitable for the thinking and problem-solving skills that students will need for future success" (Herman & Winters, 1994, p. 49). However, what we do

not know is whether performance scores represent an enduring and meaningful capability? Are they good indicators of what we think we are assessing?

Enthusiasts claimed that using an authentic assessment provided a more realistic picture of student's individual achievement and progress than a standardized test (Grace, 1992; Herman and Winters, 1994). Authentic assessment, particularly a portfolio, purportedly demonstrates growth and development over a period of time. Portfolios are collections of students' work selected by the individual students and/or teacher to represent the students' efforts, progress, and achievements over time (Gillespie, Ford, Gillespie, & Leavell, 1996). In some cases, the learners were required to assess their own work and take responsibility for their learning (Micklo, 1997). In 1988, Vermont was the first state to adopt portfolio assessment as a statewide indicator of student performance. Only Kentucky and a few other states followed the lead.

Scoring of written work in portfolios usually relies on rubrics. Supporters claim that rubrics clarify expectations, provide parameters, and offer guidance both to the teachers preparing the portfolios and to the teachers evaluating them (Burch, January 1997). Although rubrics provide some guidance, questions remain about validity. According to LeMahieu (1995b),

instability of scores may be introduced through the judgments of raters or through variability in student work, such as different prompts. It is also important that student work is not judged on factors irrelevant to the quality of the performance. Farr (1990) and Brandt (1992) claim that portfolio assessment can be considered reliable and valid. Farr (1990) addressed validity by stating that, "items selected for inclusion be labeled with the date they were written or read, the conditions under which each sample was written, and some note from the student as to his or her reaction to the work sample" (p. 103). Brandt (1992) stated that reliability could be achieved by, "knowing the behavior you're looking for and having enough evidence to feel confident that the score given is apt and representative...Second, use multiple judges where possible and require high inter-rater reliability" (pp. 35-36). The latter goes back to the previously mentioned Sung Dynasty. Between 960-1299 A.D., the Sung Dynasty introduced the practice of two independent readers for written exams.

Herman & Winters (1994) stated, "One useful approach in determining what portfolio scores mean is to look for patterns of relationships between the results of portfolio assessments and other indicators of student performance. Score meaning becomes supported when portfolio scores relate highly to other,

valued measures of the same capability...Using this approach...the researchers found moderate correlations ranging from .47 to .58 between writing portfolio scores and direct writing assessments...Similarly Gearhart and others (1993) found virtually no relationship when comparing results from writing portfolios with those from standard writing assessments. In fact, two-thirds of the students who would have been classified as "masters" based on the portfolio assessment score would not have been so classified on the basis of the standard assessment" (p. 51).

As reported by Bracey (1995), the Koretz et al study of the Vermont assessment program for the RAND Corporation found that it was hard to train large numbers of raters at a sufficient level of accuracy and that it was important to use well-standardized tasks. Vermont subsequently moved away from using portfolio scores to compare schools. Doug Walker, manager of the Vermont education department's school and instructional support team, stated: "Portfolios provide rich information about some very specific skills and knowledge, but we were concerned about their use for accountability" (Manzo, 1996, p. 3).

It was not just portfolios that gave pause to concerned psychometricians. While portfolios contain several pieces of student work, open-ended questions, as part of a single

assessment instrument, provided their own trail of doubt. Advocates of open-ended questions assert that this format eliminates the guessing or best-choice situation. It also allows students to demonstrate the process of determining a response. However Eha (1998) noted, "Proponents of multiple-choice feel that a thoughtfully constructed multiple-choice item can yield the same, or similar, results and that, for open-ended items, the measure of subjectivity introduced by adding human judges to the scoring equation-and the resulting increase in the probability of errors-is too high a price to pay for too little extra yield...At the same time there are numerous factors-such as differential application of scoring rubrics and assorted "halo effects"-that add additional error to estimates of student ability" (p. 1).

Concern about the rating of open-ended or performance items has been voiced by many experts. In 1889, F. Y. Edgeworth was the first to research essay test score validity. More recently, Herman & Winters (1994) stated, "Raters who judge student performance must agree regarding what scores should be assigned to students' work within the limits of what experts call 'measurement error...' Do they (raters) assign the same or nearly similar scores to a particular student's work? If the answer is no, then student scores are a measure of who does the

scoring rather than the quality of the work" (p. 49). LeMahieu (1995b) echoed this thought, noting concerns about the probability of obtaining acceptable levels of agreement between judges and whether the rubrics support sufficiently high expectations for students. According to Wiggins (1994), scorers tend to over-emphasize process and form criteria in scoring performance and under-emphasize or ignore impact criteria-the criteria that relate to purpose and desired effects (p. 39).

Alan Farstrup, executive director of the IRA, summarized the concerns. There is a "big issue of how quickly and how superficially should we rely on performance-based, high-stakes data when we simply don't know if it's telling us what we need. Some very serious decisions are being made about kids' lives based on instruments we've pressed into service...the consequences could be devastating" (Manzo, 2000, p. 17).

Assessment and New Jersey

New Jersey has always been in the vanguard of testing. In 1906, Henry H. Goddard, a pioneer psychometrician, established a laboratory for the study of mental deficiency at the Training School for Feeble-Minded Boys and Girls in Vineland. Later, Vineland became the site for planning the World War I Army

intelligence test (Alpha), the U.S. grandfather of the abilities testing movement.

Statewide testing began in 1972 with the first administration of the Educational Assessment Program (EAP). "The program was an outgrowth of the "Our Schools" project which established a series of outcome and process goals for New Jersey schools" (Johnson, 1985, p. 1). The primary purpose of the EAP was to assist districts in identifying program needs and providing direction for basic skills improvement. Grades four, seven and ten were tested annually; grade twelve was tested every three years.

In 1975, the New Jersey state constitution was amended to require the Legislature to establish a system of thorough and efficient education. The Legislature's response was the Public School Education Act to provide to all children in New Jersey, regardless of socioeconomic status or geographic location, the educational opportunity that would prepare them to function politically, economically and socially in a democratic society. Formally designated Chapter 212, this act became commonly known as the "Thorough and Efficient Act." Chapter 97, an amendment signed in September 1976, established uniform standards of minimum achievement in basic communication and computation

skills. This amendment is the legal basis for the use of a test as a graduation requirement in the state of New Jersey.

The test created in response to the Public School Education Act amendment was the Minimum Basic Skills Test (MBS). While the EAP identified areas of need for program improvement, the MBS was also charged with identifying student proficiency in minimum basic skills in reading and mathematics. The content of the test reflected the title. The Public School Education Act amendment further required that students who fell below the acceptable level of performance were to be provided with a program of remediation to address the identified deficiencies. As of the 1981-82 school year, ninth-grade students were required to pass the MBS in reading and mathematics as a requirement for a high school diploma.

In January 1983, New Jersey Commissioner of Education Saul Cooperman announced a new statewide testing system to replace the MBS. The commissioner's charge to the three content-area panels (reading, mathematics, writing) was to identify skills for a test more rigorous than the MBS; one that would establish higher educational standards in the state. As relayed by Tynette W. Hills (1988), chairperson of the state subcommittee on Child Development Research and Learning in HSPT Institute Focusing on Kindergarten through Third Grades, the skills and competencies

were to be those that "a student needs in order to become a productive citizen in our society" (p. 39). The first due notice administration of HSPT occurred in October 1983; the first administration as a graduation requirement for the class of 1989 was October 1985 (Hills, 1988; New Jersey Department Of Education, 1997b).

In 1988, the New Jersey Legislature passed a law that moved the High School Proficiency Test from ninth grade to eleventh grade (HSPT11). New Jersey legislators agreed that the eleventh grade test was necessary to ensure the preparation of students to participate in an increasingly complex and technological society. Since HSPT11 would be administered later in a student's high school career, it was important to identify at an earlier point those students at risk of not mastering the skills tested. Therefore, the 1988 HSPT law also established an eighth grade Early Warning Test (EWT).

Although the EWT was developed as a result of the 1988 HSPT law, the concept of early identification had a predecessor. In January 1983, the state Board of Education adopted a statewide testing system that required districts to assess their students' progress using a state-approved commercial or locally-developed test and reporting the grade-level results for students in grades three and six. "The use of tests in the early grades

serves as an "early warning system" for identifying students with skills deficiencies. To accomplish this, for each test, statewide standards were established which are equivalent to the passing scores on the High School Proficiency Test (HSPT)" (New Jersey Department of Education, 1993). In actuality, each publisher created its own standard per grade level which was then presented to the state department for approval. There is no record of any proposed standard being refused.

The 1988 eighth grade and eleventh grade skills development committees were charged with creating tests that indicated higher order thinking, reflected contemporary research in content-areas assessment and were linked to the existing ninth grade HSPT (McGettigan, 1989, 1990; New Jersey Department of Education, 1990). Increasing the connection between the two tests, the eighth grade committee was composed of educators who developed the eleventh grade skills, along with additional elementary school representatives. The final product for both tests included multiple-choice and free-response items, the latter requiring students to construct written responses.

According to New Jersey Department of Education publications of that time, HSPT11 was a rigorous test of essential skills in reading, mathematics and writing. Following three years of due-notice testing, HSPT11 was first administered

as a graduation requirement in October 1993 to all regular education eleventh-grade students (New Jersey Department of Education, 1998, 1999a).

The EWT was first administered statewide in March 1991. As with HSPT11, the New Jersey Department of Education described the EWT as "a rigorous test of essential skills in reading, mathematics and writing" (New Jersey Department Of Education, 1997b, p. 3). Further, the skills identified for the EWT were the benchmarks to the ones on the graduation test. However, the EWT was neither a graduation test nor a promotion requirement. As stated in the reports of the various skills committees, the purpose of this test was to identify students in need of remedial education services and to determine the effectiveness of the elementary curriculum in preparing students for the skills assessed by HSPT11. According to a number of documents published by the New Jersey Department of Education, students had to first master the EWT skills before mastering those on the grade eleven HSPT (New Jersey Department of Education, 1990, 1997c).

In 1992, the state Board of Education mandated the establishment and administration of a statewide fourth-grade test in New Jersey Administrative Code (N.J.A.C.) 6.8-4.6(a)1. The fourth-grade test was intended to provide an "accurate

measure of how elementary school students are progressing towards acquiring the knowledge and skills needed to graduate from high school and function politically, economically, and socially in a democratic society" (New Jersey Department of Education, 1999b, p. 5).

Although a fourth-grade test was mandated in 1992, implementation of the assessment instrument did not take place until 1997. Within this period, as part of the state's Strategic Plan for Systemic Improvement of Education, the New Jersey Department of Education engaged in a parallel task, the establishment of state academic standards (Klagholz, 1998). Adopted by the New Jersey state Board of Education in May 1996, the Core Curriculum Content Standards (CCCS) were intended to describe what all students should know and be able to do at the end of fourth grade, eighth grade, and upon completion of a New Jersey public school education.

In New Jersey, the standards took on an unusual significance. The Comprehensive Plan for Educational Improvement and Financing, published by the Department of Education in November 1995(b), recommended a new educational funding system "driven by statements that describe what students should learn as a result of having received a thorough and efficient education" (p. 25). The report further noted that the

development of the standards was designed to "assure that the funding system is predicated on high expectations for all students, rather than on minimums and basics..." (p. 25). The New Jersey standards cover eight major subject areas, mathematics, science, language arts and literacy, social studies, world languages, visual and performing arts, and comprehensive health and physical education. They also include workplace readiness expectations: career planning and work group skills; use of technology, information, and other tools; critical thinking/decision making/problem-solving; self-management; and safety principles. Through their breadth, the standards were intended to represent "the most comprehensive description ever devised of the knowledge and skills that the public education system ought to impart to all of its students" (p. 26). To further stress the importance of the standards, Commissioner Klagholz recommended that "adoption of the standards be accompanied by the modification or elimination of all competing curriculum mandates that have been accumulated over past decades in statute and regulation...and that do not reflect a systematic attempt to define the knowledge and skills all children need to thrive socially, academically and economically" (p. 26). The report further suggested that only expenditures "critically related to students' achievement of

standards" (p. 28) would be funded. A per-pupil funding target that represented a thorough and efficient education as defined by the standards would be implemented as the method of fiscally supporting education. In May, 1997, the New Jersey Supreme Court accepted this premise, linking the standards to state funding of education. This was a monumental win for Commissioner Klagholz.

The state established standards. However, a July 10, 1996 department of education release noted, "While standards point to the educational destination, they do not provide a road map for getting there...By specifying the results and not the means, the State Board has affirmed the importance of local district discretion...As a State-generated legal entity, each district exists to ensure that its students achieve, in an efficient way, the results which the standards define as a thorough public education" (Peretzman, p. 1).

New Jersey interwove the concept of standards with assessment. A May 5, 1999 memo from Commissioner Hespe stated, "The state's standardized criterion-referenced approach directly aligns the assessment program with the Core Curriculum Content Standards. In effect, the statewide assessment system establishes the performance levels for the Core Curriculum Content Standards" (p. 2).

As described by former Commissioner Klagholz (1997), "The core curriculum content standards and the related assessment program are now the sole focus of student learning in New Jersey. This new chapter of code will consolidate previous State Board rules regarding curriculum, instruction, and state assessment. It will also rescind other curriculum mandates in the code..." (p. 4).

Following the development of the standards, three new tests were introduced by the state. As would be expected in this world of accountability, the fourth-grade Elementary School Proficiency Assessment (ESPA) test specifications became aligned to the Core Curriculum Content Standards. ESPA was developed in 1996 by National Computer Systems (NCS). The first administration, May 19-29, 1997, was a practice test of the language arts literacy (reading, writing, and speaking), mathematics, and science components to determine whether the questions were at the correct level of difficulty and which questions provided the best measures of each subject. By the 2001-2002 school year, the state department of education estimated that ESPA would measure each core curriculum content standard including workplace readiness (New Jersey Department of Education 1999b, 1999c, 2000b). However, the state School Boards Association requested that ESPA be removed from the new

Standards and Assessment Code and replaced with a nationally recognized standardized test. While state board members were somewhat sympathetic, the code, adopted on April 5, 2000, continues to include ESPA (Mooney, 2000).

No scores were reported for the first ESPA administration. Scores for the second administration, May 1998, were reported in January 1999. These scores were intended solely for district use since a passing or cut-off score had, again, not been developed.

In May 1999, the ESPA was administered as an operational assessment on the mornings of May 3-7, 1999, with make-up testing on the mornings of May 10-13, 1999. In addition, a speaking administration occurred at a district-selected time period between May 10, 1999 and May 28, 1999 and science performance tasks between April 1 and June 11. Student, school, and district scores were reported for this third administration.

The stated purpose for ESPA was to assist educators in evaluating student achievement of the CCCS as well as to provide a basis for monitoring school, district and statewide implementation of the standards. The results were intended to serve as primary indicators for determining those students in need of instructional intervention and to indicate the progress students were making in mastering the knowledge and skills required by the end of fourth grade as well as the knowledge and

skills they would need to succeed on the Grade Eight Proficiency Assessment (GEPA). According to a press release by the Department of Education (Hespe, 1999c), "the statewide average test results in Mathematics showed that 40% were partially proficient, 44% proficient, and 16% were advanced proficient. Fourth grade Language Arts Literacy results were 59% partially proficient, 41% proficient and less that 1% advanced proficient. Results in Science at this grade level were 14% partially proficient, 52% proficient and 34% advanced proficient" (p. 2). Former commissioner Saul Cooperman, in an October 3, 1999, commentary in The Sunday Star-Ledger asked, "Are teachers doing a great job in science but a poor job in language arts?" He answered his own question, "The Department of Education has work to do in order to present compelling evidence as to why the science scores were so high and the language arts scores so low" (Section 10, p. 7). Obviously, the department of education felt the same way. According to Cooperman, they awarded a $90,000 contract to Keyes Martin, a public relations firm in East Hanover, New Jersey, to help convey the scores in a proper framework.

Originally entitled the new EWT the second new assessment in the state plan, GEPA, replaced the EWT in March 1999. As with ESPA, GEPA was designed to measure student and district

implementation of the Core Curriculum Content Standards. The GEPA was administered as an operational assessment on the mornings of March 9, 10 and 11, with make-up testing on March 16-18. The first administration measured language arts literacy and mathematics. As part of language arts literacy, the GEPA speaking assessment was field-tested from March 22-30, 1999. The intent was that the GEPA would be used to "indicate the progress students are making in mastering the knowledge and skills they will need to pass the HSPA" (New Jersey Department of Education, 1999b, p. 5). The same September 21, 1999 press release that revealed ESPA scores also provided information on 1999 GEPA performance, "the statewide average test results in Mathematics showed that 38% were partially proficient, 43% proficient, and 19% were advanced proficient. Eighth grade Language Arts Literacy results were 22% partially proficient, 71% proficient and 7% were advanced proficient" (p. 2). In the 2002-2003 school year, the department of education intends that GEPA will measure all core curriculum content standards including workplace readiness.

As with assessments at grades four and eight, the third piece of the planned state assessment program is in development for the eleventh-grade. The High School Proficiency Assessment (HSPA), proposed successor to HSPT11 and stated to be aligned

with the content standards, is currently being field-tested for a two-year period. According to Hespe (1999a), HSPA "will reflect curriculum content measured previously by the ESPA and GEPA." Students graduating in 2007 will be the first class required to demonstrate proficiency in all content areas including workplace readiness.

State support for assessment was highlighted in the new administrative code. According to N.J.A.C.6A:6-4.1 (2000), district boards of education are required to administer the statewide assessment system, including five major components: ESPA, GEPA, HSPA, the Special Review Assessment (SRA), and the Alternate Proficiency Assessment (APA).

As previously noted, in an interesting decision, the New Jersey Supreme Court tied the new state testing program to school funding. In the early 1970s, school funding court action began with Robinson verses Cahill. In 1973, the New Jersey Supreme Court ruled that the heavy reliance on property taxes for education discriminated against poor districts. This started a school funding battle that continued in 1981 with Abbot verses Burke. Through the advocacy of the Education Law Center, funding became a persistent legal thorn. On May 14, 1997, the state Supreme Court, in an adjunct to its order to state officials to immediately increase funding for poor urban districts, stated

that the New Jersey Core Curriculum Content Standards might eventually improve educational opportunity. "The content and performance standards define educational opportunity required by the Constitution. It is an effort that warrants judicial deference. We therefore conclude that the standards are facially adequate as a reasonable legislative definition of a constitutional thorough and efficient education" On May 21, 1998, the state Supreme Court reiterated that support for the standards.

Regardless of state Supreme Court support for the standards and assessments, not everyone agreed. A June 9, 1999 letter from the Chair of the New Jersey Association of School Administrators (NJASA) Ad Hoc Committee for Standards and Assessment Code, Dr. Susan LeGlise, to Commissioner Hespe stated, "The DOE must ensure that the Elementary School Proficiency Test (ESPA), Grade 8 Proficiency Assessment (GEPA) and the High School Proficiency Assessment (HSPA) are both valid and reliable in accordance with testing procedures of organizations such as Educational Testing Service" (p. 5). On October 21st of that year, Daniel Money, cabinet member of the New Jersey Principals and Supervisors Association (NJPSA) presented testimony to the state board on behalf of the association. The testimony urged a delay in the implementation of the testing program until there was a careful

review of reliability and validity. This was followed up in January 2000 by the New Jersey School Boards Association (NJSBA). The NJSBA directors approved a seventeen-page series of recommendations on the state testing program covering everything from administration of the tests to their evaluation. Called a "scathing report" by The Star-Ledger (January 29, 2000), the recommendations were formally presented to the state Board of Education on February 16, 2000. Despite the concerns voiced by every major, state, professional, education organization, on April 5, 2000, an unanimous Legislature adopted the new administrative code (N.J.A.C. 6A:6), entitled the Standards and Assessment Code (Mooney, 2000).

Although N.J.A.C. 6A:6 was adopted (Hespe, 2000b), some of the negativism obviously made an impact on Commissioner Hespe. In a memo to Chief School Administrators, dated April 17, 2000 and distributed through the county superintendents' April round tables, the commissioner addressed both revisions to the current ESPA and GEPA schedules and the potential assessment of the standards in visual and performing arts, health and physical education, and world languages. Commissioner Hespe made it a point to note, "The department and the educational community remains committed to the implementation of the Core Curriculum Content Standards. The successful implementation of the

Standards rest on the alignment of local curricula and the development of an aligned statewide assessment system." At the same time, he stated that, "In response to recommendations made to the department by many in the education community, two additional refinements to the assessment schedule are being implemented effective immediately" (p. 1). The identified refinements involved the move of the operational administrations of the ESPA and GEPA social studies components to the 2001-2002 school year. This was surprising since both ESPA and GEPA social studies field tests had been conducted during the 1999-2000 school year.

As for the potential assessment of the three additional standards, Commissioner Hespe wrote, "The department remains committed to these subject areas and their importance to a well-rounded, world-class education...However, a number of legitimate questions have been raised regarding the most appropriate way to test in these areas while ensuring that the goals of the Core Curriculum Content Standards are met. Assessment development committees for visual and performing arts and health and physical education have been working for over a year to grapple with the many issues involved with testing in these content areas." The commissioner further noted that, "All field test activities for the visual and performing arts, health and

physical education and world languages scheduled for the 2000-2001 school year will be postponed" (p. 2). Once again, it appears, the New Jersey state assessment program, envisioned by former Commissioner Klagholz (1998) to be complete for ESPA in 2001-2002 and complete for GEPA in 2002-2003, is riding the seesaw created by psychometric questions and inconsistency.

While the breadth of the assessment program is under review, the accuracy of scoring the open-ended questions, particularly the essay tests, has also been revisited. National Computer Systems (NCS) is the company contracted to score the New Jersey tests. In turn, NCS subcontracts the essays to Measurement Inc. of Durham, North Carolina for $4.50 per graded paper. According to an article in The Record (Glovin, 1998), their investigation showed that Measurement Inc. then hired college-educated jobbers for $7.25-$7.75 per hour to score the essays. After three days of training on the New Jersey rubric, up to seventy temporary workers, ranging from a former fighter pilot to an artist launching a gallery, graded an average of 150 papers in a seven-hour day. A $200.00 bonus was paid after 8,000 papers. In a telling interview, one four-year reader, Julian Harrison remembered the following:

> There were times I'd be reading a paper every 10 seconds. It was horrific...you could actually—I know

> this sounds very bizarre–but you could put a number on these things without actually reading the paper... (wrong grades) Either I read it too fast or I didn't recognize what the child [meant] or maybe I got impatient because the child's handwriting was very bad...Maybe some of the readers weren't careful enough, and maybe a child got a 3 instead of a 4 or a 2 instead of a 4 (p.8).

In a Sunday Star-Ledger article, former New Jersey education commissioner Klagholz acknowledged "Writing is hard to assess. I wouldn't say that they are bad tests, but the tests are evolving and the evolution isn't as good as in reading and math" (Alaya, 1999, p. 29).

An August 21, 2000 memo from Commissioner Hespe recognized that the concern about writing has not been satisfied. The memo stated, "Many of you have expressed concern about the language arts literacy section of the test and the results of your students. Based on your concerns and our own concern flowing from the student performance data, we are completing a thorough analysis of the language arts literacy section of the test for both 1999 and 2000." The memo continued, "At this time, we know that students performed at appropriate and expected levels in the selected response reading sections of the ESPA Language Arts

Literacy test...In consultation with our technical advisors, we have begun to review and analyze the open-ended responses and writing tasks in greater detail" (Hespe, 2000d).

At the same time, Commissioner Hespe's confidence in the overall assessment program was stated in that August 2000 memo, "The second year of test results underscore our confidence in the validity, reliability, and scoring of the statewide assessment system." The commissioner's confidence was not reflected throughout the state. An August 2000 article by Rimbach and Wiggins in The Record stated that, "state officials said almost 75 percent of the fourth-graders got one point or less on a grading scale of 0-4 on several of the six questions that require writing." On that same date, The Star-Ledger noted that nearly sixty percent of both 1999 and 2000 New Jersey fourth-graders had failed the ESPA language arts section. The article quoted Hespe, "When you have 40 percent of the students getting a zero on a questions, that's indication to me that need to re-evaluate it." However, in the same article, Hespe stated, "Regardless of what happens, this will be a hard test" (Mooney, 2000c).

With questions abounding, it appeared an appropriate time for re-evaluation. Inherent in the educational process is a cyclical review of programs. On April 12, 2000, the state

Department of Education announced that Achieve, Inc. a non-profit organization created following the 1996 national Education Summit, was hired to evaluate the standards and state tests. Ellen Schechter, the assistant commissioner for standards and assessment, stated, "It's important to get a sense of how well we are stacked up against other states and how well the standards are aligned with how we test them" (Mooney, 2000b, p. 25).

Although the state recognized the concern about the standards and assessments and openly entered into an evaluation process, the naysayers did not stop. On September 14, 2000, the executive committee of the New Jersey Association of School Administrators made two motions public. The first by Janet Kalafat, seconded by Wendy Schadt, unanimously supported a five-year moratorium on all state testing during which the assessments would be re-examined. The second motion by Steve Sokolow, seconded by Ronald Larkin, unanimously supported that standardized assessments replace the state tests during the five-year moratorium.

Accountability and Consequences

The New Jersey Supreme Court tied enormous accountability to the concept of standards and assessments.

Further court support for state assessments came from Texas. On January 7, 1999, U.S. District Judge Ed Prado ruled that the Texas high school graduation exam, a high-stakes test established through legislation in 1984, helped to erase educational disparities. He further noted that all students had an equal opportunity to learn the tested material. Attorneys for the Mexican American Legal Defense and Educational Fund (MALDEF), which filed the lawsuit on behalf of several minorities in 1997, had argued during the trial that about 7500 students each year did not pass the test and were denied a diploma (Lawton, 1997d). Governor George W. Bush stated that "This court decision is a victory for high standards and strong accountability" (Henry, 2000, p. 6D).

Accountability was also a key factor in President Clinton's May 1999 ESEA plan which called for requiring performance report cards at the state, district and school levels (Robelen, 1999). For some states, this was already in place. As noted previously, of the forty-eight states with assessment programs, twenty-four have high-stakes tests that high school students must pass in order to get a diploma. During 2000, forty states plan to issue report cards on schools based on their test performance, twenty-one plan to issue overall ratings for schools based on performance, and eighteen expect to have the legal authority to

close, takeover, or move staffs of failing schools (Jerald, 2000). Both presidential candidates have proposed tying federal education dollars to states' test scores (Associated Press, 2000).

Of course, with accountability comes pressure. At what may be the highest level of acceptance of personal accountability and ensuing pressure, Governor Gray Davis, in a June 1999 speech to the American Legion state convention in San Diego, vowed not to seek re-election in 2002 if student reading levels and test scores did not improve during his term (Johnston & Jacobson, 1999). For Gov. Davis, it was not just his political future that was at stake. His eye on accountability spread to the school level. In March 1999, the governor proposed ranking every school in the state based on test scores and other factors to be part of an index developed by the state education department. As of 2004, seniors would have to pass an exit exam before getting a diploma (Johnston, 1999). This emphasis on assessment has led to a variety of testing programs. For its fourth assessment program of the 1990s, California implemented the Standford Achievement Test-9th edition as a state test. As reported by Bradley (2000), at the local California level Ramon C. Cortines, interim superintendent of the Los Angeles school district, extended the concept of accountability to teachers. In March 2000, he

proposed to the school board tying the compensation for each teacher to their students' scores.

Connecting compensation to student achievement was not a completely new idea. In October of 1996, the Philadelphia school board approved a program to hold teachers more accountable for student performance. Cash awards were to be provided to successful schools. Titled the professions-responsibility program, the idea was to hold teachers more accountable for student test scores, attendance and graduation rates. The program linked teacher pay raises and performance reviews directly to student achievement and even allowed a teacher with chronically low-performing students to be fired. Although several states, notably Kentucky, had similar accountability, Philadelphia exceeded those in the degree to which it proposed to hold teachers responsible (Manzo, 1996a).

Accountability for schools was also a focus in Florida. On June 24, 1999, Florida released its first school-by-school report card for the state's 2500 schools. The scores were part of a new statewide accountability system that promised cash rewards to high-performing schools and state-financed vouchers to students attending those that failed. Florida assigned each public school an A, B, C, D, or F, based largely on how it performed on the state's predetermined standards for competency

on the reading and mathematics sections of the Florida Comprehensive Achievement Test, as well as the state writing test. Florida was the first state to assign specific letter grades to all of its schools (Sandham, 1999b). David Clark, spokesman for the Florida Teaching Profession-NEA, said, "The failing label is more stigmatizing than motivating. Standardized tests...are designed to be a diagnostic tool, not a rating system" (Sandham, 1999a, p. 18).

With puritan history, it was no surprise that Massachusetts also joined the schools accountability movement. In September 1999, the state board of education voted 7-1 to rate schools in two year cycles, identifying their overall performance and improvement on the Massachusetts Comprehensive Assessment System (MCAS). In the first two-year cycle, schools were put into one of six categories ranging from critically low to very high based on their 1998 test scores. The schools will be rated again on their 2000 test scores and evaluated on whether they met designated goals for improvement (Hoff, 1999e). In 1998, an ad hoc group of Massachusetts parents and educators joined forces as the Coalition for Authentic Reform in Education (CARE). Leaders included Theodore R. Sizer, founder of the Coalition of Essential Schools, and Alfie Kohn. CARE conducted extensive critiques of the 1998 and 1999 MCAS, concluding that both were

too long, poorly written, and full of questions that discouraged critical thinking (Lindsay, 2000). In spite of the CARE concerns, in May 2000, the Massachusetts state school board "unanimously approved regulations that will force secondary school mathematics teachers to take the exams (MCAS) as part of their recertification process if more than thirty percent of regular education students in their school fail" (Hoff, 2000c, p. 16).

In New York City, city and school officials relied on test scores for a variety of decisions including the annual review of each principal. Poor performing schools were placed on a state watch list called Schools under Registration Review (SURR). During the summer of 1999, the New York City board of education voted to shut down fourteen schools because of persistently low scores. Former schools' Chancellor Rudolph F. Crew stated that he had no plans to lower the bar. "We've lived with a system that for too long has sent more students to jail than Yale. The question is, do we have the capacity morally to stand for this, or will we decry what goals students are not yet reaching and say that it will never happen" (Hoff, 1999g).

New York may be known for cities and Colorado for open spaces. However, the leaders think alike. As of March 27, 2000, Colorado joined more than twenty states that rate their schools

based on student performance on state tests. Schools graded F will be allowed three years to improve. If no improvement is shown, the schools will be converted into a charter school (Sandham, 2000).

In June 1995, the Board of Education of Virginia published the current Standards of Learning (SOLs) for the core areas of English, mathematics, science and social studies. In the spring of 1998, state examinations were given for the first time in grades 3, 5, 8 and for certain core courses in high school. When the assessment system is complete, students will have to pass six of eleven end-of-course tests in predetermined areas before they are awarded a Virginia high school diploma. Students will be evaluated; so will schools. Those with low test scores will lose accreditation. According to the Board of Education of Virginia, looking at standards and their consequences as a threat to education systems is nonproductive and self-defeating (Bezy, 1999).

While Virginia joined many states in developing a high-stakes, graduation test, another southern state went further. Louisiana moved down the grades. In that state, test results determine whether fourth and eighth grade students can progress to the next level (Hoff, 2000b). Large, urban cities such as Chicago and New York have similar policies. In spite of parental

backlash, Delaware, Ohio and South Carolina are heading in the same direction (Robelen 2000). During a Fall 1999 assessment workshop in Philadelphia, conducted by The Center on Learning, Assessment, and School Structure (CLASS), a concern for the generalization of this early gate-keeping was voiced by many of the 1500 participants. This may be another example of the confusion cited by the Annenberg Institute for School Reform, located in Providence, Rhode Island. According to Warren Simmons, executive director, "...unfortunately, what we have in too many districts and states is test-driven reform masquerading as standards-based reform" (Olson, 2000b, p. 12).

Students and parents did not always welcome the assessments and accountability with open arms. In 1997, hundreds of protesting Michigan parents forced the state to change the test by refusing to allow their children to participate (Education Week, 1998a). Massachusetts students joined the fray in April 1999. During the second round of the Massachusetts Comprehensive Assessment System (MCAS), some students refused to take the test claiming that it was destructive and made the curriculum less flexible. At Danvers High School (Danvers), seven students took the option of reading a book in lieu of taking the test. Seventeen students at renowned Cambridge Rindge and Latin school also skipped the exam while the principal of Boston Young

Achievers' School, a pilot program, offered students the option of not taking the MCAS. Student protesters also wrote and phoned state representatives, picketed the state Capitol, and sent faxes to news organizations (White, 1999). During that same period, students at a top-performing school in Chicago deliberately flunked a state exam as a protest against testing frenzy (Lindsay, 2000). In California, Ohio, and Wisconsin parents also conducted grassroots campaigns to keep their children home on test days (Lindsay, 2000; Olson, 2000b).

A variety of problems across the nation with both assessment instruments and test scores substantiated the concerns of parents and students (Bradley 1997, 1999). In February 1999, New York education commissioner Richard P. Mills formed a six-member, blue-ribbon panel to determine whether a series of miscues undermined the validity of the fourth-grade reading test (Hoff, 1999a). The test was developed by CTB/McGraw-Hill.

In 1997, Kentucky state education officials decided that the results on one portion of the 1996 test would not count in the final score. Ann M. Sheadel, the chief hearing officer for the attorney general's office said that schools should not be held accountable for the mistakes and problems in an assessment system. The mistake involved an open-ended, project-based,

performance question. To Kentucky's credit, state officials conducted a reliability analysis and, discovering that scores were inconsistent, proposed eliminating the item (Jacobson, 1997). The state also replaced the contractor Advanced Systems in Measurement and Evaluation. The new test, CATS, will include more standardized tests to allow national comparisons as well as to improve reliability and validity (Kearns et al, 1999; Jacobson, 1999).

Advanced Systems in Measurement and Evaluation also had to adjust test scores in Maine for the 1995-1996 school year. The Dover, New Hampshire firm discovered that it had used an incorrect equation to calculate scores on the Maine Educational Assessment. As a result, traditionally high-achieving schools appeared to perform at a lower level than expected (Education Week, 1996).

Advanced Systems in Measurement and Evaluation is a relatively small company specializing in state assessments. CTB/McGraw-Hill and Harcourt Educational Measurement, two major forces in assessment, demonstrated that large test publishers also can have problems. Mistakes on the CTB Terra Nova, administered in New York City during the 1998-1999 school year, ranged from printing math scores on forms for the reading test to scoring problems that caused thousands of students to be

mistakenly sent to summer school or held back a grade (Viadero, 1999a, 1999b, 1999c). CTB/McGraw-Hill also notified educators in Indiana, Nevada, South Carolina and Wisconsin that the percentile rankings on the Terra Nova may be incorrect (Viadero and Blair, 1999). For its part, Harcourt agreed to give the Vermont Department of Education a package of refunds and discounts over the next two years as a penalty for scoring errors on the 1998 and 1999 state tests in mathematics and language arts. The Texas publisher miscalculated eighth grade writing scores and incorrectly listed the mathematics scores for thirty-nine high school students. Harcourt also had to rescore the 1998 fourth and eighth grade writing sections for both Vermont and Rhode Island (Bowman, 2000). As for New Jersey, the rush to develop tests led to occasional problems. An April 20, 2000 FAX, entitled Errata, from National Computer Systems to state districts listed eleven errors in the ESPA Examiner's Manual. Some districts, on Spring Break, did not receive or distribute the FAX until after the first day of testing had been completed.

Regardless of inaccuracies on specific tests or with scoring, pressure still accompanies accountability. That pressure does not just impact students and parents; it can also foster unacceptable side effects with professional staff

members. Before the June 8, 2000 Associated Press caption, "Educators cheat in pupil tests," was designed, before the Education Week caption, "As stakes rise, definition of cheating blurs," was conceived (Hoff, 2000d), problems occurred. The 1996 president of the National Association of Elementary Schools Principals, Carole Kennedy, stated that accountability was definitely a factor in the accusations of school teachers and administrators tampering with standardized tests. Stephen Klein, a senior research scientist at the RAND Corporation, provided the following 1999 quote, "There's no testing program I know of that's immune to the problem. We spend a lot of money on these tests, and it's like throwing money away because you can't ensure the validity of the test scores" (Sandham, 1999a, p. 20). In Kentucky, more than one hundred schools were investigated for cheating. In Barker, NY, the superintendent resigned and the elementary school principal was put on paid suspension in response to allegations of test tampering. In the fall of 1996, a Chicago curriculum coordinator and a school principal were put on unpaid suspension for allowing students at Clay Elementary School to practice on the Iowa Test of Basic Skills that was to be administered in the Spring of 1996. That test was the first implementation of the Chicago school board policy requiring third, sixth and eighth graders to score at a specified level on

the Iowa test before promotion to the next grade (Lawton, 1996c). Chicago scored again in the blatant category when George Schmidt, a public school English teacher and editor of the Substance newspaper, was suspended without pay for publishing test questions from the Chicago Academic Standards Exams in his paper (Boser, 2000).

Edward F. Stancik, special commissioner of investigation for New York City, stated that investigators uncovered cheating in thirty-two of the city's 675 elementary and middle schools. The investigation found that teachers and principals routinely prepped students for the exams, openly prompted them to give correct answers, and even completed portions of exams. Fifty-one principals, teachers, and aides were cited in the report. The report further suggested that the scores of up to 1,000 students were tainted by cheating adults (Hoff, 1999g).

A similar response to pressure created problems in Texas. The Texas Assessment of Academic Skills (TAAS), developed in 1993 and first administered in 1994, dominates spring throughout the state (Johnston, 1998). In March 1999, investigators found that Houston students were given oral prompting, that answer keys were used to correct student answers, and that test security was lacking. In Austin, school officials are charged with changing student data to raise state accountability

ratings. On a sixteen count indictment, Deputy Superintendent Kay Psencik was accused of modifying student test data and of failing to stop members of her staff from making changes. She was not alone. The test contractor identified an additional ten districts as having an excessive number of erasure corrections over a three-year period (Johnston, 1999; Johnston and Galley, 1999). Continuing the problem, in April 2000, a grand jury in Austin indicted eighteen school officials for allegedly altering student tests (Associated Press, 2000).

Rhode Island may be smaller than New York or Texas, but the testing irregularities are also strong. Teachers in some schools kept past copies of state tests to help prepare students. According to Commissioner of Education Peter J. McWalters, "It became clear that the scope of the breach was excessive" (Archer, 1999, p. 28). An interesting point here is that the investigation found that educators appeared unaware that they were compromising the integrity of the exam.

Psychometric Considerations

Being unaware falls into many categories. With all the maneuvering in the assessment world to determine accountability, there appears to be little concern or awareness of the need for psychometric accuracy. W. James Popham (1999), emeritus

professor in the graduate school of education and information studies at the University of California, Los Angeles, has stated that educators do not know much about measurement. In the mid-1960s, when standardized achievement tests were first used to satisfy the program evaluation requirements of the Elementary and Secondary Education Act (ESEA) of 1965, educators sat back because they did not know better. In the 1980s, when newspapers began to rank schools according to students' scores, again educators sat back because they did not know better. "As the misuse of standardized-test scores became ever more pervasive, educators continued to assent because they weren't sufficiently knowledgeable. Well, that simply must change" (p. 32).

The lack of awareness is inconsistent with the more than hundred year history of psychometrics. In the late 1800s, an interest in psychometrics came to the fore. In 1864, the Reverend George Fisher introduced the first criterion-referenced tests. In 1889, F. Y. Edgeworth researched essay test score validity (Nitko, 1983). Francis Galton was the first person to carry out a systematic examination of a school population. He developed methods of mathematical analysis that are the basis of contemporary testing statistics. Galton's work was continued by mathematician/biologist Karl Pearson at University College,

London. Aside from developing tests, Pearson gave his name to a correlation statistic (Smith, 1986).

While psychometrics is not a new field, those educated in the discipline are relatively small in number. The terms standardized, reliable, and valid are recognized by the general public. However, not many understand the implications. This, undoubtedly, led to the previously noted statement by Stake, "it appears that many states have not taken adequate steps to validate their assessment instruments, and that proper studies would reveal important weaknesses" (Heubert and Hauser, 1998, p. 179).

Several states and districts have made a conscious effort to ensure the strength of the recognized assessment. Vermont was one state to take appropriate steps in reviewing the state assessment. Researchers Daniel Koretz, Brian Stecher, Stephen Klein and Daniel McCaffrey of the RAND Corporation specifically focused on performance assessment questions. As reported in 1994, their study found that it was hard to train large numbers of raters at a sufficient level of accuracy and that it was important to use well-standardized tasks. Their report closed by urging ongoing evaluation in our national experimentation with innovative large-scale performance assessments as a tool of educational reform.

Virginia was another state to take a proactive stance toward reviewing a new assessment. As reported by Portner (1999), the Standards of Learning tests, first administered in spring of 1998, were analyzed by testing experts from Michigan State University, the University of Virginia, and Virginia Commonwealth University. Results of a comparison of the assessments and the Standards of Learning led to a conclusion that the tests had content validity. The panel also compared passing scores for students in grades 3, 5, and 8 with the students' latest scores on the Stanford Achievement Test-9th Edition and Virginia's Literacy Passport Test. According to Cameron Harris, Virginia's assistant superintendent for assessment, the conclusion was that the tests were not "widely different assessments" (p. 3). The panel further concluded that the tests were fair and reliable because students tended to perform consistently, whether passing or failing, throughout different parts of the tests.

Similar research was recently conducted in New York by test publisher, ERB (2000). The study compared performance on state-mandated fourth grade and eighth grade Language Arts and Mathematics assessments to that of the same students on similar tests of the CTP III. Focusing on one suburban district with slightly over 130 students per grade level, the publisher

obtained simple correlations of scaled score performance and compared the rank orders created by the two tests. The data produced statistically significant results. The correlation for eighth-grade Language Arts was a strong .77 while that for eighth-grade Mathematics was a very high .86. The fourth-grade results were not as high with a Language Arts correlation of .61 and a Mathematics correlation of .63.

On a national basis, the review of NAEP called broad attention to problems in the testing world. In 1998, NRC, the same group that reviewed NAEP, sounded a warning about high-stakes testing, especially when a single test is used for graduation or retention. This concurred with Secretary of Education Richard W. Riley's statement that test results should not be used for high stakes decisions unless they have been validated for this purpose (Manzo, 1997).

Summary

Testing has been a force since 1122 B.C. Loved, scorned, cherished, villified, the word, test, has meant assessment, achievement, acceptance, accountability. In the international, national and statewide arenas, tests have transitioned from measuring basic skills to measuring standards. Throughout this transition, the concept of accountability has strengthened. With

rewards and sanctions for students, schools, teachers, and even a governor at stake, it is essential that the assessments are based on good indicators of the content area to be measured and that they are strong enough to support the assigned accountability. From its inception, ETS had the right focus - psychometric responsibility. Regardless of the state, the nation, the test, the most important question should be - is the assessment valid?

Chapter III

RESEARCH METHODOLOGY

Introduction

The purpose of this chapter is to describe the data population, provide background on the instruments used to gather the data, identify the procedures for the collection and treatment of the data, and present the study design for each of the four major research questions.

Population

The data were developed in a suburban, Monmouth County, New Jersey public school district. The district is a member of District Factor Group (DFG) I. The largest DFG, I includes 105 of the state's 573 districts. This suburban district has over three thousand students, providing an excellent sampling population. Written permission to conduct the study was received from the interim superintendent on February 18, 2000.

The district design locates all classes of a grade level in the same building. This reduces variability in the implementation of curriculum and instruction. Student achievement is strong. The district was recognized by Governor Christine Todd Whitman in her 1996 State of the State address as providing an outstanding education at less than the average per

pupil cost, thereby validating the curriculum and achievement levels already in place. In addition, the high school has repeatedly been identified by New Jersey Monthly as one of the top in the state (DeMonte and Rapp, 1998; Nusser and Faris, 2000).

The district provides a traditional education with regular K-12 articulation among faculty and administrators. An approved curriculum for each content area and grade level creates scope and sequence guidelines. While there is individualization of instruction, grade-level and department meetings extend articulation and ensure the consistency of curriculum implementation. At the high school level, there are sixteen honors courses and eleven Advanced Placement courses. No currently popular education reform (block scheduling, etc.) appears as a variable in this study.

Eighty-six percent (184) of the two hundred and fifteen members of the high school class of 1999 took both the PSAT/NMSQT and the HSPT11 in October 1997. The PSAT mean verbal score was 54.7; the PSAT mean mathematics score was 59.2. Based on PSAT/NMSQT results, nine students were named National Merit Semi-finalists and twenty-two students were named National Merit Commended Scholars. This is within the norm for the district, again supporting the implemented curriculum.

Ninety-six percent (207) of the class of 1999 took the SAT and the HSPT11. For these students, the SAT verbal mean was 561; the SAT mathematics mean was 613; the total SAT mean was 1174. Eighty-one percent of the class of 1999 selected a four-year college including Columbia (8), Cornell (2), Emory (1), Georgetown (2), Johns Hopkins (2), Princeton (3), Rutgers (27), Seton Hall (1), United States Military Academy (1), United States Naval Academy (1), and the University of Pennsylvania (5). Fourteen percent selected a two-year college; five percent of the class decided on employment. This data again supported the fact that the implemented curriculum and instruction prepared the student body for future choices.

Growth was a factor in the district. In 1995, there were 2773 students; in 1999, the district numbered 3236 students. At the same time, mobility, as determined by state records, was relatively low averaging around four percent, consistent with the average mobility index for DFG I districts. This allowed a strong population for the comparison of EWT and HSPT11 scores. Eighty percent of the class of 1999 took both tests while residing in the district. The low mobility rate was also a factor in allowing the comparison of the scores of over two hundred and thirty students on the 1999 GEPA with their scores

on a national, standardized, achievement test administered in 1998.

An April 1, 1998 memo from then education commissioner Leo Klagholz to the state board of education noted that the state no longer required districts to continue to administer assessments other that those in the state testing program. However, the district board of education opted to require a national, standardized, achievement test for fourth grade students. This provided the opportunity to correlate the scores of over two hundred and fifty students on the 1999 ESPA with their scores on the 1999 administration of a national, standardized test.

Instruments

The instruments in this study included four tests of the New Jersey state assessment program and three nationally administered, standardized tests. The state assessment instruments were HSPT11, EWT, GEPA and ESPA. Each was recognized as the official state test for the grade level at the time of administration. The national, standardized tests were the SAT, Preliminary Scholastic Assessment Test (PSAT), and the Comprehensive Testing Program, third edition (CTP III), all developed under the auspices of ETS. Only the reading, writing

and mathematics components of each test were included in the study.

New Jersey Assessments

HSPT11: In 1988, the New Jersey Legislature passed a law (18A:7C-6) that moved the High School Proficiency Test from ninth grade to the eleventh grade. HSPT11 has served as a graduation requirement for all New Jersey public school students who entered ninth-grade or adult high school on or after September 1, 1991. Following three years of due-notice testing, HSPT11 was first administered as a graduation requirement in October 1993 to all eleventh grade students except special education students whose individual education plan exempted them from the requirement.

HSPT11 consists of three tests: reading, mathematics and writing. The reading test measures literal and inferential comprehension through four types of passages: narrative text, informational text, persuasive/argumentative text, workplace text. The passages are selected from published books, newspapers, magazines, and government and business papers (New Jersey Department of Education, 1985, 1990, 1997e).

The mathematics test requires students to solve problems of basic mathematics, algebra and geometry. Specifically, the test

measures numerical operations, measurement and geometry, patterns and functions, data analysis, and fundamentals of algebra.

The writing test requires students to read passages and answer multiple-choice questions that measure a student's revising and editing skills such as correct usage, sentence construction, and organization. This test also includes a writing task where students construct meaning by writing an essay.

To determine the passing score on HSPT11, a three-day standard-setting study was conducted in December 1993, using information from the October 1993 administration. The intent of the study was to "describe and delineate the level of performance indicative of eleventh-grade minimal mastery performance and to establish a score which would differentiate minimal mastery performance from non-mastery performance in each subject" (New Jersey Department of Education, 2000a, p. 5). Participants in the study included educators from secondary schools and higher education, students, parents, and representatives from business. For multiple-choice items, a modified Angoff approach to standard-setting was used. For the open-ended items, judges decided which of a group of exemplar responses would pass or fail. Following the standard-setting

study, passing scores for each section of the test were set in January 1994.

Through statistical equating which employed a common-item equating design, each subsequent HSPT11 was linked to the October 1993 HSPT11. According to the department of education (2000a), "Equating is a statistical procedure that converts test scores from different test forms to the same score scale. The HSPT11 employs a common-item equating design. This design utilizes information derived from a set of items which appear in common on all test forms to be equated...Once the performance of the different groups on the common items is calculated, equating methodology is used to place raw scores from each test form on the HSPT11 score scale" (p. 5). All post-1993 tests were, and are, linked through equating to ensure that the HSPT11 levels of difficulty do not change from one test administration to another. This purportedly allowed test results to be compared from year to year because they represented equal levels of achievement based upon the October 1993 HSPT11 administration.

The total HSPT11 reading score is determined through a combination of multiple-choice and open-ended items. The total HSPT11 mathematics test score is based on a combination of multiple-choice, open-ended, grid-response, and graphical-response formats. Open-ended items require students to construct

their own written responses rather than choosing an option; grid-response items require students to use grids to code their numeric or symbolic responses; graphical response formats require students to graph solutions on a prepared and designated grid. The total HSPT11 writing score is based on a combination of correct multiple-choice items and the number of points received for the writing task. The writing task is weighted to account for sixty percent of the total score.

HSPT11 scores are reported as scale scores with a range of 100 to 500. The passing score is 300. Possible points are calculated by assigning each multiple-choice and grid-response item one point and each open-ended item up to three points. According to the New Jersey Department of Education (2000a), "the points received for an open-ended item are based on rater evaluation. For each item, the number of points received is the average of the number of points given by two raters" (p. 6). The number of points for the writing task ranges from 2 through 12. "Each essay is rated by professional readers using a rating scale that ranges from 1 (inadequate command of written language) to 6 (superior command of written language). The number of points a student can receive for the Writing Task ranges from 2 to 12...Each essay is rated by two separate readers. If the number of points given by both readers is the

same or within one point, the ratings are added together and the student receives the total number of points from the two readers" (pp. 6-7).

Students have four opportunities to pass each section of HSPT11. Those who do not pass all sections by twelfth grade and have completed, or will complete within the appropriate timeframe, all other high school graduation requirements, may be eligible to demonstrate their mastery of the required skills through the Special Review Assessment (SRA). In the SRA process, a team of educators examines other evidence of mastery and determines whether the student has the skills to achieve the equivalent of a passing score on HSPT11. Students who do not pass the SRA or the HSPT11 do not receive a high school diploma. That student may return for another year of school and retake the HSPT11; return to school only for HSPT11 testing; enroll in an adult high school and take the HSPT11 there; or take and pass the General Educational Development (GED) test.

This study included data from the October 1997 HSPT11. According to the New Jersey Department of Education (1999e), 64,058 New Jersey eleventh grade students participated in that administration. Of these, 11,774 were members of DFG "I" districts.

EWT: The Early Warning Test measured eighth-grade skills that indicated progress toward mastery of the essential skills that are tested on HSPT11 (New Jersey Department of Education, 1995a, 1997a). Consisting of three sections - reading, mathematics and writing - the EWT was first administered in March 1991. An Eighth-Grade Reading Skills Development Committee, consisting of teachers and administrators, identified reading skills to serve as appropriate benchmarks to the HSPT11 skills. The committee determined that both groups should be expected to comprehend narrative text, informational text, and persuasive/argumentative text. The eleventh-grade workplace text was changed to everyday text for the eighth-grade population. As with the eleventh-grade, literal and inferential comprehension were measured through reading the lines, reading between the lines and reading beyond the lines. Open-ended items requiring constructed written responses accompanied the normal multiple-choice format.

The Eighth-Grade Mathematics Skills Committee also patterned the EWT after the HSPT11. Items measured numerical operations, measurement and geometry, patterns and relationships, data analysis, and pre-algebra. Multiple-choice, open-ended, grid-response, and graphical-response formats were included.

In a similar fashion, the Eighth-Grade Writing Skills Development Committee followed the HSPT11, but at a lower level of complexity and sophistication. The multiple-choice portion of the test included reconstructing meaning by revising/editing the written text of another writer. Subclusters included sentence mechanics, based on common mechanical errors; sentence construction, which involved selecting a revision that corrected an error in sentence construction; precision and coherence in which students selected words, phrases, etc. that completed a partially constructed sentence in written text; sentence combining; transitions and logical progressions; and focus and organization, where students organized the content of written text.

A three-day standard-setting study was conducted in May 1994, using information from the March 1994 administration. Participants in the study included eighth-grade teachers. For multiple-choice items, a modified Angoff approach to standard-setting was used. For the open-ended items, judges decided which of a group of exemplar responses fell into three categories: does not need instructional intervention, may or may not need instructional intervention, needs instructional intervention. Through statistical equating which employed a common-item equating design, each subsequent EWT was linked to the March

1994 EWT to ensure that levels of difficulty were comparable. March 1994 was selected as the base year to coincide with the first group of students taking the HSPT11.

EWT scores were reported as scale scores with a range of E001-E250. The E250 was a theoretical ceiling; E indicated an EWT score as opposed to the HSPT11 score. However, the two scales overlapped numerically. EWT scores were reported as Proficiency Level I (E150-E250), clearly proficient; Proficiency Level II (E100-E149), minimally competent and may or may not need instructional intervention; and Proficiency Level III (E001-E099), those who need instructional intervention (New Jersey Department of Education, 1997d).

This study includes data from the March 1995 EWT. According to the New Jersey Department of Education (1995a), 71,219 regular education eighth grade students took at least one section of the 1995 EWT.

GEPA: Originally entitled the New EWT, the Grade Eight Proficiency Assessment was introduced in March 1999. When complete, it will measure language arts literacy, mathematics, social studies, visual and performing arts, science, health/physical education, world languages and workplace readiness skills. The first administration included language

arts literacy and mathematics. For consistency in this study, only these two tests were included.

According to the New Jersey Department of Education, the language arts literacy assessment measures students' achievement in reading and writing. There are four content clusters: writing, reading, working with text, and analyzing/critiquing text. The writing cluster consists of three activities: a writing/speculate task in response to a picture, a persuasive writing task, and a passage that students edit and revise. The 1999 language arts literacy test also included a speaking component. This was one of the performance assessments that the state intended to develop and administer collaboratively with districts and charter schools. Since the performance assessments were scored locally using a state-developed scoring rubric, reliability was difficult to ensure. The state did not include speaking in the 2000 administration Therefore, the speaking component is not included in this study.

The mathematics test measures knowledge and skills in four content clusters: number sense, concepts and applications; spatial sense and geometry; data analysis, probability and discrete mathematics; patterns, functions and algebra (New Jersey Department of Education, 1999b).

The 1999 GEPA included both machine-scored, multiple-choice items and rater-scored, open-ended items. Between June 8-11, 1999, the proficiency levels were determined by panelists who represented the various DFGs and regions of the state. The panelists were either practicing teachers or curriculum supervisors in one of the content areas. A holistic classification method was used for the proficiency-level setting study. Based on thirty-three student test booklets, pre-selected by the testing company to cover the range of student performance, the panelists individually classified each booklet as partially proficient, proficient, or advanced proficient. Partially proficient students were considered to be below the state minimum level of proficiency, needing additional instructional support. The method was holistic because judges had to consider responses to multiple-choice, short constructed-response, and open-ended questions. The panelists next received twenty-two booklets around the preliminary "proficient" cut score and twenty-two around the preliminary "advanced proficient" cut score. This second set of booklets promoted discussion and allowed for score adjustments. Although only 77 booklets were involved in the standard setting, it should be noted that 76,390 general education students took at least one

section of the March 1999 test. Data from this administration was included in the study.

Using a logistic regression method, two cut-off scores were calculated based on the judges' classifications. The final "cut" scores were based on a total of seventy-seven booklets. The two scores yielded three proficiency levels. Through a statistical equating procedure, the GEPA will be comparable from year to year; the March 1999 GEPA serves as the base year (New Jersey Department of Education, 1999b).

The maximum 1999 GEPA score was 300. Students who scored between 250 and 300 were rated advanced proficient; those who scored between 200 and 249 were considered proficient; and those who scored below 200 were identified as partially proficient. According to the New Jersey Department of Education, these terms are specified in federal law and nationally accepted. The term partial notes that all students have some degree of proficiency in a content area and can build upon that base (New Jersey Department of Education, 1999b, 1999c 1999g).

In all subject areas, the 1999 scores were reported as scale scores. An October 22, 1999 memo from the assistant commissioner stated that eight percent of the students across the state scored advanced proficient on the 1999 GEPA Language Arts Literacy test, seventy-seven percent scored proficient, and

fifteen percent scored partially proficient. On the Mathematics test, twenty-two percent scored advanced proficient, forty-six percent scored proficient, and thirty-two percent scored partially proficient.

ESPA: In 1996, the fourth grade Elementary School Proficiency Assessment was developed by National Computer Systems in response to N.J.A.C. 6.8-4.6(a)1. The first administration was May 1997. After two pilot administrations, each with problems in setting scores for interpretation, the first official administration occurred in May 1999. When complete, ESPA will measure language arts literacy, mathematics, social studies, visual and performing arts, science, health/physical education, world languages and workplace readiness skills. The 1999 administration included language arts literacy with a speaking component, mathematics, and science with a laboratory component. The speaking and science laboratory aspects have been removed for the 2000 testing. Scored in-district, there was a question of consistency across interpretations. For this study, only the state-scored language arts literacy and mathematics tests are included.

According to the New Jersey Department of Education, the language arts literacy assessment measures students' achievement in reading and writing. There are four content clusters:

writing, reading, working with text, analyzing/critiquing text. The writing cluster consists of two activities: a writing/speculate task in response to a picture and a writing/analyze task that relates to a poem. The mathematics test measures knowledge and skills in five content clusters: number sense, operations and properties; measurement; spatial sense and geometry; data analysis, probability and discrete mathematics; patterns and algebra (New Jersey Department of Education, 1999b).

The 1999 ESPA included both machine-scored, multiple-choice and rater-scored, open-ended items. Between August 10-13, 1999, the proficiency levels were determined by panelists who represented the various DFGs and regions of the state. As with GEPA, the panelists were either practicing teachers or curriculum supervisors in one of the content areas. A holistic classification method was used for the proficiency-level setting study. Based on thirty-three student test booklets, covering the range of student performance as determined by the testing company, the panelists individually classified each booklet as partially proficient, proficient or advanced proficient. The method was holistic because judges had to consider responses to multiple-choice, short constructed-response, and open-ended questions. The panelists next received twenty-two booklets

around the preliminary "proficient" cut score and twenty-two around the preliminary "advanced proficient" cut score. This second set of booklets promoted discussion and allowed for score adjustments. As with GEPA, the final "cut" scores were based on a total of seventy-seven booklets.

Using a logistic regression method, two cut-off scores were calculated based on the judges' classifications. The two scores yielded three proficiency levels. The maximum score was 300. Students who scored between 250 and 300 were rated advanced proficient; those who scored between 200 and 249 were considered proficient; and those who scored below 200 were identified as partially proficient. (New Jersey Department of Education, 1999d). In all subject areas, the scores were reported as scale scores (New Jersey Department of Education, 1999f).

This study includes data from the May 1999 ESPA. According to the New Jersey Department of Education (1999f), 87,471 general education fourth grade students took at least one section of that test. An October 22, 1999 memo from the assistant commissioner stated that one percent of the students scored advanced proficient on the ESPA 1999 Language Arts Literacy test, forty-five percent scored proficient, and fifty-four percent, over half the fourth-graders in the state, scored partially proficient. On the Mathematics test, eighteen percent

scored advanced proficient, forty-seven percent scored proficient, and thirty-four percent scored partially proficient.

Nationally Administered Standardized Tests

Two of the nationally administered, standardized tests in this study are developed by ETS. The third was designed by the Educational Records Bureau (ERB); the operations for this test are handled by ETS.

SAT: Originally named the Scholastic Aptitude Test, briefly called the Student Assessment Test (1994), the SAT is now known officially by its initials. Regardless of title, the test was designed in 1926 by psychometrician Carl Campbell Brigham as an exam that Ivy League schools could use to award scholarships to students who did not come from elite New England families. Currently, the College Board describes it as a test of developed mathematics and verbal reasoning skills.

The SAT was selected for this study because of its impact on education. According to Lemann (1999), "by 1990, much of the curriculum in American elementary and secondary education had been reverse-engineered to raise SAT scores...Average SAT scores were widely used as a measure of school quality...to the taker it was a scientific, numeric assignment of worth..." (p. 273). Schwartz (1999) called it the single, most important test for

high school students. Further support for the test came from the New Jersey Department of Education which used SAT scores as one of three criteria to determine funding levels in 103 high achieving districts (New Jersey Principals and Supervisors Association, 1998a).

SAT I is a three hour, primarily multiple-choice test, that measures verbal and mathematical reasoning abilities. The verbal test includes reading passages, sentences, and word pairs. The mathematics test centers on problems involving arithmetic, algebra and geometry, the latter two based on a year of algebra and some geometry course work.

To determine the score on the SAT, a raw score is calculated with each correct answer receiving one point. To correct for random quessing on multiple-choice questions a fraction of a point is subtracted. One-fourth is deducted for five multiple-choice answers; one-third for four multiple-choice answers; one-half for three multiple-choice answers. No points are deducted for student-produced mathematics responses. Raw scores are then converted to scaled scores. Through an equating process, scaled scores are adjusted to compensate for small variations in difficulty from one edition to another. Equating ensures that students' scores are not impacted by the edition of the test nor by the abilities of the group taking the test.

Total scores are reported on a 200-800 scale. If the student does not answer any verbal or mathematics questions, that student receives a 200 on the respective test. A perfect score on each exam is 800.

In 1999, the year of the study, a record two million students took the SAT. The national math average fell one point to 511; the national verbal average stayed at 505 for the fourth consecutive year. The total national mean score was 1016. As noted above, in the studied district, the 1999 SAT verbal mean was 561; the 1999 SAT mathematics mean was 613; the total 1999 SAT mean was 1174.

PSAT: The Preliminary Scholastic Assessment Test (PSAT) was originally developed as a practice SAT. Following the lead of the SAT, the PSAT was initially named the Preliminary Scholastic Aptitude Test and is now known as the Preliminary Scholastic Assessment Test or by the initials, PSAT. Until 1997, the test replicated the SAT, including only mathematics and verbal questions. In 1997, a writing skills component was added following an agreement with the U. S. Department of Education office for civil rights to settle a gender-discrimination complaint by the Cambridge, Massachusetts watchdog group, FairTest (Lawton, 1996; Manzo, 1998a).

The verbal test measures sentence completion, analogies and critical readings. The latter includes a variety of passages of variable length, from 400-850 words on diverse subject matter including humanities, natural sciences, and social sciences. For each passage, a brief introduction orients students to the selection. Critical reading questions measure vocabulary in context, comprehension and extended reasoning.

The mathematics test includes problem solving and quantitative comparisons; it measures student's ability to apply mathematical concepts and interpret data. In preparation, students need a basic knowledge of arithmetic, algebra, and geometry. Students may use any four-function, scientific, or graphing calculator on the test. Questions include multiple-choice and constructed-response (grid) formats.

The multiple-choice writing section measures the identification of sentence errors, improvement of sentences and improvement of paragraphs. ETS states that students who have written and revised essays in school have undoubtedly addressed the writing problems addressed in this test.

Verbal, mathematics and writing sections are scored on a 20-80 scale. Adding a zero transforms the score into an SAT score. In October 1997, the Preliminary Scholastic Assessment Test was administered to more than a million students. Across

the country, the average PSAT scores of juniors tend to be between 47 and 49. In the studied district, the 1997 PSAT verbal mean was 54.1; the 1997 PSAT mathematics mean was 59.2.

The PSAT is the qualifying test for the National Merit Scholarships. Prior to the introduction of a writing skills section, the verbal score was doubled and added to the mathematics score for the NMSQT index. Beginning with the October 1997 test, the NMSQT selection index was determined from the sum of the verbal, mathematics and writing skills scores. The selection index for NMSQT varies from state to state. The New Jersey index is always among the most rigorous (Chiles, 1998); New Jersey is third in the nation in participation, behind Connecticut and the District of Columbia (Heyboer, 1998). For the 1997 PSAT/NMSQT, the New Jersey selection index for commended scholar was 199; the New Jersey selection index for semi-finalist was 219. Commended scholars are those who score in the top five percent for a particular state; semi-finalists are those who score in the top half of one percent in a given state. Based on 1997 PSAT/NMSQT results, nine students in the studied district were named National Merit Semi-finalists and twenty-two students were named National Merit Commended Scholars.

CTP III: The Comprehensive Testing Program (CTP) was developed by the Educational Testing Service (ETS) for the

Educational Records Bureau (ERB). ERB was founded in 1927 by Dr. Ben Wood, Chair of the Department of Statistics at Columbia University, "to develop educational assessment instruments capable of measuring achievement and ability along the whole continuum of student capability and performance. At the time, Dr. Wood's concern was that available tests did not provide adequate measures for high ability, high performing students in the independent and suburban schools of the country" (Educational Testing Services, 1995, p. 1). For this study, the current and third edition, CTP III, was used.

In grades four through eight, CTP III includes the following tests: verbal ability, reading comprehension, vocabulary, writing mechanics, writing process, quantitative ability and mathematics. For the purposes of this study, verbal ability, reading comprehension, quantitative ability and mathematics were administered to fourth grade students while verbal ability, reading comprehension, writing mechanics, writing process, quantitative ability and mathematics were included in the study of the 1999 eighth grade class.

According to the Comprehensive Testing Program III Technical Report (Educational Testing Service, 1995), the identified CTP III tests provide information on the following concepts:

- Verbal Ability measures the ability to appropriately apply knowledge of printed language structure and meaning, to use cognitive strategies in analyzing information and drawing inferences, to deduce and generalize verbal attributes and to predict outcomes and evaluate the appropriateness of predictions and strategies.
- Reading Comprehension measures a student's general comprehension of written material through vocabulary, the recall of explicit information, the identification of main ideas plus the abilities to hypothesize and to summarize using information from a passage.
- Writing Mechanics measures a student's understanding of writing conventions such as punctuation, capitalization, language use and spelling.
- Writing Process measures the understanding of the processes and components of effective composition.
- Quantitative Ability measures the ability to apply knowledge of mathematical concepts and principles.
- Mathematics measures a student's application of mathematical knowledge to solve problems, compute solutions, reason, estimate and communicate in accordance

with the National Council of Teachers of Mathematics (NCTM) Standards for Curriculum and Evaluation.

CTP III content validity was determined by matching the items on the proposed test with the curriculum of the user schools. Member schools included both public and non-public participants with the emphasis on high-achieving schools in each category such as Dalton (New York City), Pingry (New Jersey), Peck (New Jersey), and Pennington (New Jersey) schools. Therefore, the content-area curriculum could be assumed to be rigorous.

The concurrent, external aspect of construct validity was determined by correlating scores on CTP III tests with scores for the same population on corresponding National Achievement Tests (NAT) and Developing Cognitive Abilities Tests (DCAT), both published by the American Council on Testing (ACT), the largest psychometric competitor of ETS. The NAT series was considered to be comparable to the CTP III achievement tests while the DCAT was considered comparable to the CTP III verbal and quantitative abilities tests.

For this study, data from the April 26-30, 1999 administration of CTP III, Level D was included for the fourth grade analysis while data from the April 27-May 1, 1998

administration of CTP III, Level E was included for the GEPA analysis.

Data Collection Procedures

In the studied district, test scores are maintained on a data base. For the purposes of this study, the following scores for the high school class of 1999 were reviewed: SAT scores, provided by ETS; October 1997 PSAT scores, provided by ETS; October 1997 HSPT11 scores, provided by the New Jersey Department of Education; March 1995 EWT scores, provided by the New Jersey Department of Education.

The 1999 eighth grade class was also studied. The following scores were reviewed: March 1999 GEPA scores, provided by the New Jersey Department of Education; April 1998 CTP III, Level E scores, provided by ETS/ERB.

Finally, the 1999 fourth grade class was studied. The following scores were reviewed: May 1999 ESPA scores, provided by the New Jersey Department of Education; April 1999 CTP III, Level D scores, provided by ETS/ERB.

Data Analysis Procedures

This construct validity study focused on the external component. According to Messick (1993), validation is scientific

inquiry into the "degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on test scores" (p. 1). In situations emphasizing accountability, whether selection, placement, graduation, certification and/or program evaluation, the external aspect of validity is the preferred format.

A test is valid if it measures what it is supposed to measure; the conclusions are accurate and useful; and there are no constant errors. The external component of validity measures the extent to which performance on a test is related to external variables. For example, two tests of mathematics computation at the same grade level should show a strong correlation. According to Messick (1993), the external component involves correlations between the total scores and/or subscores of the two tests. Herman and Winters (1994) report that correlations between traditional, multiple-choice tests and tests including, or primarily, performance based is supported through the Gearhart et al (1993) study of portfolios and the work of Koretz et al (1993, 1994) in Vermont.

Correlations of the total test scores were based on scale scores. According to Petersen, Kolen and Hoover (1989), "The process of associating numbers with the performance of examinees is referred to as scaling, and this process results in a score

scale...A score scale refers to numbers, assigned to individuals on the basis of test performance, that are intended to reflect increasing levels of achievement or ability" (p. 221). Scale scores are usually, but do not have to be, linearly related to raw scores. In this study, all scale scores are linear. Therefore, although the scales are different, the scale scores can be compared since each scale reflects linear, increasing levels of achievement.

For the subtests of the 1999 ESPA and GEPA, scale scores were not available. Therefore, raw scores were used. This was the first year for the GEPA test and was the first official administration of the ESPA test. According to Petersen, Kolen and Hoover (1989), "the raw score scale on an initial form of a test can be used as the primary score scale, and scores on subsequent test forms can be equated to the scale of the initial form" (p. 222).

Correlation describes the degree of relationship between two or more variables. That relationship is reported as a correlation coefficient, a number between -1.00 and 1.00. For this study, the correlations were conducted using the Pearson formula. Named for British scientist, Karl Pearson, this correlation coefficient describes the linear relationship between pairs of quantitative variables.

The Pearson product-moment correlation, also known as the Pearson correlation coefficient and the Pearson r, is the most widely used form of correlation measures. According to Cohen, Swerdlik and Smith (1992), it is the statistical tool of choice when the relationship is linear and when the variables are continuous, that is they can theoretically take any value. "The formula for Pearson r takes into account the relativity of each test score's position" (p. 118). Pearson r was selected for this study because the identified scale scores and raw scores reflect linear, increasing levels of achievement. In addition, the relativity of each score's position in the distribution as a whole was the concept being studied as opposed to the score as an inherent value. Interpretation of a Pearson correlation further notes that the sign of r indicates the type of linear relationship, positive or negative, while the value of r, without regard to the sign, indicates the strength of the relationship (Witte, 1993).

The value of r can be further interpreted by deriving the coefficient of determination or r^2, also written as R^2. According to Witte (1993), the r^2 indicates the "proportion or percent of a perfect relationship" by describing the "proportion of explained or predictable variability" (p. 153). Moreover, $100(1-r^2)$

indicates the variance that can be attributed to chance, error, or other unexplained factors (Cohen, Swerdlik, Smith, 1992).

Two approaches were incorporated in the study. Concurrent, external, construct validity was established when r=.7 or higher. According to Fordham University lectures by Dr. Anastasi (1963), a recognized psychometrics expert, the acceptable coefficient for validity for standardized assessment(s) was r=.7 or higher. A review of validity coefficients in technical manuals for national tests supported this premise. Moreover, at r=.707, the coefficient of determination and $100(1-r^2)$ are equivalent with half of the variability explained and half due to chance, error, or other unexplained factors. Once the correlation is below r=.7, the majority of the variability is due to unexplained or unpredicted factors. In lieu of other research on the topic, r=.7 has been identified as the significant number. In instances where r fell between .60 and .69, the data were considered to indicate moderate validity and the need for further study. Therefore, each null hypothesis (H_{01} - H_{030}) was rejected if r=.6 or higher. With correlations between r=.0 and r=.59, the tests were considered not to be valid. This qualified as a decision rule. According to Witte (1993), "a decision rule specifies precisely when H_0 should be rejected" (p. 225).

While the correlations between HSPT11, PSAT and SAT, as well as those between GEPA, ESPA and CTP III, all measured the concurrent, external aspect of construct validity, the correlations between the EWT and HSPT11 looked at predictive validity. According to Anastasi (1982), predictive validity is required when there is interest in predicting or determining the relationship between two measures over an extended period of time. Messick (1993) describes it as "the extent to which an individual's future level on the criterion is predicted from prior test performance" (p. 16). To further support the findings in this area, regression analyses were also conducted on the EWT/HSPT11 data. Through regression analysis, data are used to identify relationships among variables. The relationships can then provide a basis for predictions. This study included linear regression which measures the constant rate of increase of one variable with respect to another (Glasserman, 1999; Levine et al, 1999).

The following study design was implemented:

Research Problem 1

Validity of HSPT11 through a Correlation of Performance with SAT and PSAT Scores

TEST	1998	1997
SAT	Verbal Mathematics	
PSAT		Verbal Mathematics Writing
HSPT11		Verbal Mathematics Writing Essay

Research Problem 2:

Predictive Value of EWT in Determining HSPT11 Performance

TEST	1997	1995
HSPT11	Reading Mathematics Writing Essay	
EWT		Reading Mathematics Writing Essay

Research Problem 3

Relationship between Performance on GEPA and CTP III, Level E

TEST	1999	1998
GEPA	LA Literacy Reading Writing Mathematics	
CTP III Level E		V. Ability Reading W. Process Q. Ability Mathematics

Research Problem 4

Relationship between Performance on ESPA and CTP III, Level D

TEST	1999
ESPA	LA Literacy Reading Writing • Poem • Picture Mathematics
CTP III Level D	V. Ability Reading Q. Ability Mathematics

Chapter IV

RESULTS

Introduction

This chapter presents an overview of the validity study, provides a statistical analysis of the quantitative data related to each of the four research questions, and concludes with a summary of the findings.

Overview

This was a validity study of the New Jersey state testing program. According to Messick (1993), validation is scientific inquiry into the "degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on test scores" (p. 1). In determining validity of high-stakes tests, those that hold accountability for students and districts, the external aspect is the preferred format (Messick, 1993).

The study examined the concurrent, external aspect of construct validity of three tests in the New Jersey state assessment program: Elementary School Proficiency Assessment (ESPA), Grade Eight Proficiency Assessment (GEPA) and High School Proficiency Test 11 (HSPT11). Using a Pearson r, correlations were conducted on scores obtained by the same

students on both the relevant state test and on a grade and content appropriate, national, standardized test. The three recognized, national, standardized tests in this study are published under the auspices of ETS: SAT, Preliminary Scholastic Assessment Test (PSAT), Comprehensive Testing Program third edition (CTP III).

The Pearson r is the most widely used form of correlation measures. According to Cohen, Swerdlik and Smith (1992), it is the statistical tool of choice when the relationship is linear and when the variables are continuous. Theoretically, continuous variables can take any value. The Pearson r was selected for this study because the identified scale scores and raw scores reflected linear, increasing levels of achievement. In addition, the study reviewed the relativity of the position of each score in the distribution. The inherent value of the score was not a factor.

The study also examined the predictive validity of the EWT by correlating the scores for the same students on both the 1995 EWT and the 1997 HSPT11, the latter taken three years later in junior year of high school. This followed the preferred process for predictive validity described by both Anastasi (1982) and Cronbach (1984). Again, the Pearson r was used.

To aid in understanding the strength of the relationships, the coefficients of determination, r^2, were calculated. R^2 describes the "proportion of explained or predictable variability" existing between two variables (Witte, 1993, p. 153). Moreover, $100(1-r^2)$ indicates the variance that can be attributed to chance, error, or other unexplained factors (Cohen, Swerdlik, Smith, 1992).

Two approaches were incorporated in the study. Concurrent, external, construct validity was established when r=.7 or higher. According to Fordham University lectures by Dr. Anastasi (1963), a recognized psychometrics expert, the acceptable coefficient for validity for standardized assessment(s) was r=.7 or higher. A review of validity coefficients in technical manuals for national tests supported this premise. Moreover, at r=.707, the coefficient of determination and $100(1-r^2)$ are equivalent with half of the variability explained and half due to chance, error, or other unexplained factors. Once the correlation is below r=.7, the majority of the variability is due to unexplained or unpredicted factors. In lieu of other research on the topic, r=.7 has been identified as the significant number. In instances where r fell between .60 and .69, the data were considered to indicate moderate validity and the need for further study. Therefore, each null hypothesis (H_{01}

- H_{030}) was rejected if r=.6 or higher. With correlations between r=.0 and r=.59, tests were considered not to be valid. This qualified as a decision rule. According to Witte (1993), "a decision rule specifies precisely when H_0 should be rejected" (p. 225).

The data were developed in a suburban, New Jersey community. Eighty-six percent of the two hundred and fifteen members of the high school class of 1999 took both the PSAT/NMSQT and the HSPT11 in October 1997. Ninety-six percent of the class of 1999 took the SAT and the HSPT11. The study was restricted to scores for those one hundred and sixty students who took all four tests: EWT, HSPT11, PSAT, and SAT.

Approximately eighty percent of the class of 1999 took both the EWT and the HSPT11 while residing in the district. All one hundred and seventy-three students who remained in the district between eighth and eleventh grades were included in the study.

The low mobility rate was also a factor in allowing the comparison of the scores of two hundred and thirty-two students on the 1999 GEPA with their scores on a national, standardized achievement test administered in 1998. In addition, a board of education decision to administer a national, standardized test along with ESPA provided the opportunity to correlate the scores of two hundred and fifty-three fourth-grade students on the 1999

ESPA with their scores on the 1999 administration of a national, standardized test.

Findings

Research Question 1

A. Determine the concurrent, external, construct validity of the HSPT11 using the PSAT as an external criterion measure

The first research question was to determine if the HSPT11 was a valid test with sufficient weight to warrant the use as a graduation requirement. It was recognized that the HSPT11 will be replaced by the High School Proficiency Assessment (HSPA), currently being field-tested. However, both tests claim to measure what should be learned at the completion of thirteen years of schooling; both were constructed, or are being constructed, by state department of education selected committees using the same format; and the development of each was designed under the guidance of the same state department employee. Therefore, since no data are available for the HSPA, this study sought to establish the validity of HSPT11. The information provided will be important in determining the future progress of the in-development HSPA.

To determine external validity, the October 1997 HSPT was correlated to the performance of the same students on the

October 1997 PSAT and a 1998 SAT. According to Schwartz (1999), the SAT has become the "single most important test for American high-school students - an academic and psychic rite of passage that strongly influences future educational options..." (p. 30). The Preliminary Scholastic Assessment Test (PSAT) was originally developed as a practice SAT. Until 1997, the test replicated the SAT, including only mathematics and verbal questions. In 1997, the composition changed with the addition of a writing skills component. The PSAT and SAT are further connected by having the same ETS managers for the verbal and mathematics components.

The following five null hypotheses were used to test the concurrent, external, construct validity of HSPT11:

- H_{01}: The correlation between the 1997 HSPT11 Reading test and the 1997 PSAT Verbal test is between 0 and .59.
- H_{02}: The correlation between the 1997 HSPT11 Writing test and the 1997 PSAT Verbal test is between 0 and .59.
- H_{03}: The correlation between the 1997 HSPT11 Writing test and the 1997 PSAT Writing test is between 0 and .59.
- H_{04}: The correlation between the 1997 HSPT11 Essay test and the 1997 PSAT Writing test is between 0 and .59.
- H_{05}: The correlation between the 1997 HSPT11 Mathematics test and the 1997 PSAT Mathematics test is between 0 and .59.

Part A of the first research question measured the validity of the HSPT11 as determined through a correlation with the PSAT. Data in Table 1 were used to capture H_{01} - H_{05}.

Table 1

CLASS of 1999

1997 HSPT11 / 1997 PSAT Correlations

HSPT11	PSAT	N	Pearson r	r^2
Reading	Verbal	160	.71	.50
Writing	Verbal	160	.51	.26
Writing	Writing	160	.49	.24
Essay	Writing	160	.33	.11
Mathematics	Mathematics	160	.72	.52

The 1997 HSPT11 and the 1997 PSAT were completed by the data population within a two-week interval. Due to the short period of time, a strong correlation between tests would be expected. However, the data did not support this premise. The correlations in Table 1 ranged from a low of .33 to a high of .72.

Two null hypotheses were rejected, thereby implying that validity was established for two HSPT11 tests based on PSAT criteria. Null hypothesis H_{05}, which stated that the correlation between the HSPT11 Mathematics test and the 1997 PSAT Mathematics test was between 0 and .59, was rejected. The correlation between the two tests was .72, indicating a strong,

positive relationship between the tests. The r^2, a direct measure of the strength of the relationship, was .52. This identified that the two tests had slightly more than half of a perfect relationship, 52% of explained variability.

Null hypothesis H_{01}, which stated that the correlation between the HSPT11 Reading test and the 1997 PSAT Verbal test was between 0 and .59, was also rejected. The correlation between the two tests was .71, again indicating a strong, positive relationship between the two tests. The r^2 was .50. From the perspective of r^2, a correlation of .71 is half as strong as a perfect relationship of 1.00 with 50% of the variance explained by the two tests. Based on the data, both the HSPT11 Mathematics test and the HSPT11 Reading test were valid as demonstrated through the concurrent, external aspect of construct validity.

Testing of the three null hypotheses in the area of writing produced different results. Null hypothesis H_{02}, which stated that the correlation between the HSPT11 Writing test and the 1997 PSAT Verbal test was between 0 and .59, was supported. The correlation between the two tests was .51. While this correlation was positive, it could not be interpreted as strong. That fact was reinforced by the r^2 of .26, identifying a relatively weak relationship. Following a similar pattern was

null hypothesis H_{03}, which stated that the correlation between the HSPT11 Writing test and the 1997 PSAT Writing test was between 0 and .59. For H_{03}, the correlation between the two tests was .49; the r^2 was .24. From the perspective of the r^2 for the correlations which tested both H_{02} and H_{03}, in each instance it identified a relationship that was half as strong as the relationship demonstrated in the correlations of the Mathematics and Reading tests, described above in H_{05} and H_{01}. The weakness of the correlations in H_{02} and H_{03} was further supported by the fact that approximately three-quarters (74%, 76%) of the variance in each could be attributed to chance, error, or other unexplained factors.

As with H_{02} and H_{03}, the third writing null hypothesis, H_{04}, which stated that the correlation between the HSPT11 Essay test and the 1997 PSAT Verbal test was between 0 and .59, was also supported. The correlation between the two tests was a low, weak .33. This correlation was so low that the abilities tested would have to be independent. The r^2 was .11, indicating that only 11% of the variance in student performance overlapped between the two measures.

Comparing the coefficients of determination for H_{02}, H_{03}, and H_{04}, it should be noted that the relationships in H_{02} and H_{03}, while unsatisfactory, were twice as strong as that between the

HSPT11 Essay test and the PSAT Verbal test, highlighting the weakness of the Essay component. Moreover, the relationships noted through the correlations of the Mathematics and Reading tests (H_{05} and H_{01}) were almost five times as strong as that of the Essay test.

Based on the data for H_{02}, H_{03}, and H_{04}, neither the HSPT11 Writing test as a total nor the Writing Task (Essay) component were valid tests as demonstrated through the concurrent, external aspect of construct validity. The very low correlation and coefficient of determination raised particular concerns about the appropriateness of the Writing Task (Essay) as an assessment instrument that holds accountability for districts and students.

B. Determine the concurrent, external, construct validity of the HSPT11 using the SAT as an external criterion measure

The following three null hypotheses were used to test the concurrent, external aspect of construct validity of the HSPT11 as determined through a Pearson r between the eleventh grade test and the SAT:

- H_{06}: The correlation between the 1997 HSPT11 Reading test and a 1998 SAT Verbal test is between 0 and .59.
- H_{07}: The correlation between the 1997 HSPT11 Writing test and a 1998 SAT Verbal test is between 0 and .59.

- H_{08}: The correlation between the 1997 HSPT11 Mathematics test and a 1998 SAT Mathematics test is between 0 and .59.

Part B of the first research question measured the validity of the October 1997 HSPT11 as determined through a correlation with a 1998 SAT. As previously noted, Schwartz (1999) stated that the SAT has become the "single most important test for American high-school students - an academic and psychic rite of passage that strongly influences future educational options..." (p. 30). Data in Table 2 were used to capture H_{06} - H_{08}.

Table 2

CLASS of 1999

1997 HSPT11 / 1998 SAT Correlations

HSPT11	SAT	N	Pearson r	r^2
Reading	Verbal	160	.77	.59
Writing	Verbal	160	.50	.25
Mathematics	Mathematics	160	.72	.52

The data in Table 2 supported the findings listed in Table 1 for each of the HSPT11 tests. Two null hypotheses were rejected, thereby implying that validity was established for two HSPT11 tests based on SAT criteria. Null hypothesis H_{06}, which stated that the correlation between the 1997 HSPT11 Reading test and a 1998 SAT Verbal test was between 0 and .59, was rejected. The correlation between the two tests was .77, indicating a very

strong, positive relationship between the tests. A more definitive correlation than that in Table 1 (.71), the data strengthened the conclusion that the HSPT11 Reading test was valid as demonstrated through the concurrent, external aspect of construct validity. The HSPT11 Reading test was, therefore, a good measure of the content area. In addition, the r^2 for H_{06}, a direct measure of the strength of the relationship, was .59, identifying a relationship that was approximately 60% of a perfect 1.00 with 59% of the variability explained.

Following in the same direction, null hypothesis H_{08}, which stated that the correlation between the 1997 HSPT11 Mathematics test and a 1998 SAT Mathematics test was between 0 and .59, was also rejected. The correlation between the two tests was a strong .72, the same number identified in Table 1 for the correlation between the HSPT11 Mathematics test and the PSAT Mathematics test. In each case, the r^2 was .52, indicating that over half of the variance could be explained by the two tests. As with reading, the data in the two tables demonstrated that the Mathematics test was valid as demonstrated through the concurrent, external aspect of construct validity, leading to the conclusion that the assessment was an effective measure of that discipline.

Recognizing that the PSAT and SAT are connected, with the same managers for the verbal and mathematics components, the parallel performance was expected. However, since the PSAT was administered within two weeks of HSPT11, while the SAT was administered at a point within one year of HSPT11, higher correlations would be expected between HSPT11 and PSAT. Therefore, the unexpected, slightly higher correlations between the HSPT11 Reading test and its SAT counterpart must be noted.

Table 2 also identified a performance parallel to that in Table 1 for the HSPT11 Writing test. The null hypothesis was supported for H_{O7} with a correlation of .50 between the HSPT11 Writing and an SAT Verbal test. This number is consistent with the two Table 1 correlations for the Writing test, H_{O3} (.49) and H_{O4} (.51). All three represent positive correlations; however, none of the three is strong. The r^2 in Table 2 for Writing was .25, indicating that only one-quarter of the variance could be explained by the two tests with three-quarters (75%) of the variance attributed to chance, error, or external factors. The r^2 further denoted that the correlation for the Writing tests described a relationship that was less than half as strong as the relationships for the Reading and Mathematics tests. Each r^2 in Table 2 was consistent with the data reported in Table 1. Based on the data in both tables, it was apparent that the

HSPT11 Writing test was not valid as demonstrated through the concurrent, external aspect of construct validity.

All correlations for Research Question 1 (A and B) were significant at the .01 level.

Research Question 2

Determine the predictive validity of the EWT using the HSPT11 as the criterion measure

The second research question was to determine if the EWT did predict performance (predictive validity) on the HSPT11, laying a foundation for questioning if the GEPA can predict performance on the HSPA and, as an extension, if the fourth-grade ESPA can predict performance on the eighth-grade GEPA. According to Messick (1993), the construct validity framework allows a rational basis for prediction. "It leads us to address, as well, varieties of discriminant evidence essential in the construct validation of both predictor and criterion measures (p. 77). Quoted by Messick, Guion (1976) stated that the predictive hypothesis was "the outcome of a rational process linking the domain theory to the choice of criterion and predictor constructs as well as to the empirically grounded construct interpretations of the criterion and predictor measures." He continues that what is appraised in predictive validity is the "validity of the hypothesis of a relationship between the test

and a criterion measure" (p. 77). In perhaps more direct words, Cronbach (1984) described the need to compare a test to the prediction for a "straightforward empirical check on the value of the test for predictive validity" (p.103). According to Anastasi (1982), predictive validity is required when there is interest in predicting or determining the relationship between two measures over an extended period of time. "If we want to use test scores to predict outcome in some future situation, such as an applicant's performance in college, we must use tests with high predictive validity against the specific criterion" (p. 30). As with concurrent validity, the two sets of data must always be on the same individual.

Predictive validity answers the challenge of the EWT, and potentially the GEPA and ESPA. The skills identified for the EWT were the benchmarks for those on HSPT11. In Department of Education meetings throughout New Jersey, it was stated that the test should be used to identify students who might have difficulty passing the high-stakes HSPT11. As with the EWT and HSPT11, the same relationship is expected between the GEPA and the in development HSPA as well as between the ESPA and the GEPA. According to Gronlund (1981), "If the results are to be used to predict student success in some future activity, we

should like them to provide as accurate an estimate of future success as possible" (p. 65).

The following four null hypotheses were used to test the predictive validity of the EWT as determined through a Pearson r between the eighth grade test and the eleventh grade HSPT11:

- H_{09}: The correlation between the 1995 EWT Reading test and the 1997 HSPT11 Reading test is between 0 and .59.
- H_{010}: The correlation between the 1995 EWT Writing test and the 1997 HSPT11 Writing test is between 0 and .59.
- H_{011}: The correlation between the March 1995 EWT Writing Task (Essay) component and the October 1997 HSPT11 Writing Task (Essay) component is between 0 and .59.
- H_{012}: The correlation between the 1995 EWT Mathematics test and the 1997 HSPT11 Mathematics test is between 0 and .59.

The second research question measured the external, predictive validity of the 1995 EWT as determined through a correlation with the 1997 HSPT11. Data in Table 3 were used to capture H_{09} - H_{012}.

Table 3

CLASS of 1999

1995 EWT / 1997 HSPT11 Correlations

EWT	HSPT11	N	Pearson r	r^2
Reading	Reading	173	.67	.45
Writing	Writing	173	.47	.22
Essay	Essay	173	.22	.05
Mathematics	Mathematics	173	.79	.62

The correlations in Table 3 ranged from a low of .22 to a high of .79. One null hypotheses was rejected, thereby implying that predictive validity was established for one EWT test based on HSPT11 criteria. Null hypothesis H_{012}, which stated that the correlation between the 1995 EWT Mathematics test and the 1997 HSPT11 Mathematics test was between 0 and .59, was rejected. The correlation between the two tests was .79, indicating a very strong, positive relationship. The r^2, a direct measure of the strength of the relationship, was .62. This identified a relationship approximately two-thirds that of a perfect 1.00 with 62% of the variability subject to prediction. From Table 3, it is apparent that the EWT Mathematics test was a good predictor of performance on the HSPT11 Mathematics test and an appropriate indicator for student and district accountability. This conclusion was strengthened through a regression analysis that indicated that a 1997 EWT Mathematics passing score of 100

was equivalent to a 1999 HSPT11 Mathematics passing score of 336. The regression analysis supported the observation that the EWT Mathematics test served as a good indicator for predicting success on the HSPT11.

Null hypothesis H_{09}, which stated that the correlation between the 1995 EWT Reading test and the 1997 HSPT11 Reading test was between 0 and .59, was rejected. With a correlation of .67, this null hypothesis fell into the area that indicated moderate, rather than definitive, validity. Although this null hypothesis was referred for further review as per the study descriptors, a regression analysis strengthened the probability that the 1997 EWT Reading test did not predict performance on the 1999 HSPT11 Reading test. According to the regression analysis on the Class of 1999 population, a minimal EWT Reading passing score of 100 led to a failing score of 171.79 on the 1999 HSPT11 Reading. Therefore, even with moderate predictive validity, the EWT Reading test was not an effective predictor for passing or failing the similar test of the HSPT11.

Null hypothesis H_{010}, which stated that the correlation between the 1995 EWT Writing test and the 1997 HSPT11 Writing test was between 0 and .59, was supported. The correlation between the two tests was .47. While positive, this number represented only a moderate relationship. According to the r^2,

less than one-quarter of the variability in HSPT11 could be predicted from the EWT. From the perspective of the r^2, the Pearson r identified a relationship that was slightly less than one-third of the relationship demonstrated in the correlation of the Mathematics tests. This weakness was further supported by the fact that over three-quarters (78%) of the variance could be attributed to chance, error, or other unexplained factors. Based on the correlation, it can be stated that the EWT Writing test did not demonstrate predictive validity.

The correlation that supported H_{010} (.47) was similar to the data on the HSPT11 Writing test listed in Table 1 and Table 2. While H_{02}, H_{03} and H_{07} tested the concurrent, external aspect of construct validity rather than predictive validity, the closeness of the correlations on the Writing test must be noted. H_{02} was supported with a Pearson r of .51; H_{03} was supported with a Pearson r of .49; and H_{07} was supported with a Pearson r of .50. The consistency of the four Writing correlations strengthened the fact that the New Jersey Writing tests are not valid. This raised questions as to appropriateness of the tests both as accountable assessment instruments and as measures of the content area.

A regression analysis on the tests in H_{010} led to another interesting dimension for writing. According to the regression

analysis on the Class of 1999 population, a minimal EWT Writing passing score of 100 led to a passing score of 336 on HSPT11, similar to the regression for the mathematics test. This result further demonstrated the instability of the writing test. Writing is the last communication skill to develop. In general, students read before they write. Therefore, the reading scores should be stronger than the writing scores. This is consistent with the predictive validity correlations. The regression analysis led to further questions about the use of the state writing assessment as an indicator of future performance.

The questions raised in Table 1 about the Essay test were intensified by the data presented in Table 3 for H_{011}. Null hypothesis H_{011}, which stated that the correlation between the 1995 EWT Essay test and the 1997 HSPT11 Essay test was between 0 and .59, was supported. The correlation between the two tests was a very low .22. The correlation was so low that the abilities tested would have to be independent. The r^2, which provides a direct measure of the strength of a relationship, was .05. This indicated that only 5% of the variability in the HSPT11 Writing Task (Essay) could be predicted from the EWT Writing Task (Essay). From the same data, it was further discerned that 95% of the variance in this correlation was due to chance, error, or other factors, an extremely high

percentage. Based on the correlation, it can be stated that the EWT Essay test did not demonstrate predictive validity.

As with the Writing test as a whole, the correlation that supported H_{011} (.22) was similar to the data on the HSPT11 Essay test listed in Table 1. While H_{04} tested the concurrent, external aspect of construct validity rather than predictive validity, the closeness of the correlations on the Essay tests must be noted. H_{04} was supported with a Pearson r of .33; H_{011} was supported with a Pearson r of .22. The consistency of the two, extremely low correlations strengthened the fact that the New Jersey Essay tests did not demonstrate validity and need a close review before continued implementation or replication.

Based on the correlations, the EWT Mathematics test demonstrated predictive validity with the HSPT11 Mathematics test, thereby indicating that the EWT Mathematics test predicted performance on the HSPT11 Mathematics test. The EWT Reading test fell just outside the range of predictive validity but close enough to warrant further study. However, a regression analysis supported the probability that the 1997 EWT Reading test was not a good predictor of a passing or failing score on the 1999 HSPT11 Reading.

Despite having the same core of item developers and a consistent objective format, neither the EWT total Writing test

nor the Writing Task (Essay) component demonstrated predictive validity with aligned tests on the HSPT11. Particularly disappointing was the correlation between the two essay tests since they were designed under the same format and scored with the same rubric under the auspices of the same scoring company. This is important to note since the EWT held accountability for students and districts. It is also important to note since the New Jersey Department of Education has announced the same predictive role for the new ESPA and GEPA tests, again incurring accountability for students and districts.

All correlations for Research Question 2 were significant at the .01 level.

Research Question 3

Determine the concurrent, external, construct validity of the GEPA using the CTP III as an external criterion measure

The third research question was to determine if the eighth-grade GEPA, first administered in March 1999, was valid as determined through correlation to a recognized, independently validated, national, standardized test, taken by the same students within a one-year time period and measuring the same content areas. Nine null hypotheses were used to test the concurrent, external aspect of construct validity of the GEPA as

determined through a Pearson r between the eighth grade test, administered in March 1999, and the CTP III, Level E test, administered in April 1998:

- H_{O13}: The correlation between the March 1999 GEPA Language Arts Literacy test and the April 1998 CTP III, Level E Verbal Ability test is between 0 and .59.
- H_{O14}: The correlation between the March 1999 GEPA Language Arts Literacy test and the April 1998 CTP III, Level E Reading Comprehension test is between 0 and .59.
- H_{O15}: The correlation between the March 1999 GEPA Language Arts Literacy test and the April 1998 CTP III, Level E Writing Process test is between 0 and .59.
- H_{O16}: The correlation between the March 1999 GEPA Reading subtest and the April 1998 CTP III, Level E Verbal Ability test is between 0 and .59.
- H_{O17}: The correlation between the March 1999 GEPA Reading subtest and the April 1998 CTP III, Level E Reading Comprehension test is between 0 and .59.
- H_{O18}: The correlation between the March 1999 GEPA Writing subtest and the April 1998 CTP III, Level E Verbal Ability test is between 0 and .59.

- H_{o19}: The correlation between the March 1999 GEPA Writing subtest and the April 1998 CTP III, Level E Writing Process test is between 0 and .59.
- H_{o20}: The correlation between the March 1999 GEPA Mathematics test and the April 1998 CTP III, Level E Quantitative Ability test is between 0 and .59.
- H_{o21}: The correlation between the March 1999 GEPA Mathematics test and the April 1998 CTP III, Level E Mathematics test is between 0 and .59.

The third research question measured the concurrent, external aspect of construct validity of the March 1999 GEPA as determined through a correlation with the April 1998, CTP III, Level E. Data in Table 4 were used to capture H_{o13} - H_{o21}.

Table 4

Eighth Grade Class of 1999

April 1998 CTP III, Level E / March 1999 GEPA Correlations

GEPA	CTP III	N	Pearson r	r^2
LA Literacy	Verbal Ability	232	0.55	.30
LA Literacy	Reading Comprehension	232	0.49	.24
LA Literacy	Writing Process	232	0.59	.35
Reading	Verbal Ability	232	0.53	.28
Reading	Reading Comprehension	232	0.47	.22
Writing	Verbal Ability	232	0.41	.17
Writing	Writing Process	232	0.48	.23
Mathematics	Quantitative Ability	232	0.83	.69
Mathematics	Mathematics	232	0.82	.67

The correlations in Table 4 ranged from a low of .41 to a high of .83. Two null hypotheses were rejected, both impacting the Mathematics test, thereby implying that validity was established for one GEPA test based on CTP III criteria. With similar correlations, null hypotheses H_{o20}, which stated that the correlation between the March 1999 GEPA Mathematics test and the April 1998 CTP III, Level E, Quantitative Ability test was between 0 and .59, and H_{o21}, which stated that the correlation between the March 1999 GEPA Mathematics test and the April 1998

CTP III, Level E, Mathematics test was between 0 and .59, were both rejected. The correlation between the two tests in H_{O20} was .83; the correlation between the two tests in H_{O21} was .82. Both correlations indicated an extremely strong and positive relationship. The r^2 for H_{O20}, a direct measure of the strength of the relationship, was .69; the r^2 for H_{O21} was .67. In each case the relationship was more than two-thirds that of a perfect 1.00. Based on the data in Table 4, the GEPA Mathematics test was valid as demonstrated through the concurrent, external aspect of construct validity. Therefore, the GEPA Mathematics test was an appropriate measure of the discipline.

The null hypotheses were supported for H_{O13} - H_{O19} with correlations ranging from .41 to .59. Three of the correlations fell in the .50 to .60 range. H_{O15}, which stated that the correlation between the March 1999 GEPA Language Arts Literacy test and the April 1998 CTP III, Level E, Writing Process test was between 0 and .59, was supported with a Pearson r of .59. While the correlation was positive, it was not strong enough to demonstrate validity. The r^2 was .35, indicating a relationship between the two tests that was only about one-half as strong as the relationships identified with the Mathematics test in H_{O20} and H_{O21}. H_{O13}, which stated that the correlation between the March 1999 GEPA Language Arts Literacy test and the April 1998 CTP

III, Level E, Verbal Ability test was between 0 and .59, was also supported. The correlation was .55; the r^2 was .30. Again, the correlation was positive but not sufficiently strong while the coefficient of determination identified a relationship between the tests that was less than half that shown through the Mathematics tests. In addition, H_{O16}, which stated that the correlation between the March 1999 GEPA Reading subtest and the April 1998 CTP III, Level E Verbal Ability test was between 0 and .59, was supported. The correlation was .53; the r^2 was .28. From the r^2, it was apparent that slightly more than one-quarter (28%) of the variance was due to the two tests while almost three-quarters (72%) of the variance could be attributed to chance, error, or external factors.

The remaining four correlations fell in the .40 to .50 range. Three of the four were close, ranging from .47 to .49. Null hypothesis H_{O14}, which stated that the correlation between the March 1999 GEPA Language Arts Literacy test and the April 1998 CTP III, Level E, Reading Comprehension test was between 0 and .59, was supported with a correlation of .49; H_{O19}, which stated that the correlation between the March 1999 GEPA Writing subtest and the April 1998 CTP III, Level E, Writing Process test was between 0 and .59 was supported with a correlation of .48; and H_{O17}, which stated that the correlation between the

March 1999 GEPA Reading subtest and the April 1998 CTP III, Level E, Reading Comprehension test was between 0 and .59 was further supported with a correlation of .47. In each case, the r^2 (H_{O14}: $r^2 = .24$; H_{O19}: $r^2 = .23$; H_{O17}: $r^2 = .22$) indicated that each pair of tests had a relationship less than one-quarter that of a perfect correlation, while over three-quarters of the variance in each case could be attributed to chance, error, or external factors.

The lowest correlation was found for H_{O18} which stated that the correlation between the March 1999 GEPA Writing subtest and the April 1998 CTP III, Level E, Verbal Ability test was between 0 and .59. This null hypothesis was supported with a correlation of .41. The r^2 was .17 indicating that only 17% of the variability could be explained while 83% of the variance could be attributed to chance, error, or external factors. The r^2 further denoted that the correlation described a relationship that was approximately one-quarter as strong as the relationship for the Mathematics tests in H_{O20} and H_{O21}.

Based on the data, the GEPA Language Arts Literacy test and subtests were not valid as demonstrated through the concurrent, external aspect of construct validity. This raised a question as to the appropriateness of their use, especially since the GEPA has accountability issues for students and districts. Placing

students in a basic skills Language Arts program as a result of GEPA scores needs to be reviewed. Of particular note was the fact that the GEPA Reading test had low correlations of .47 and .53, indicating an area of concern not raised with the EWT.

It should be noted that the population was large for a correlation study and that the students in the data population had a consistent, written, board-approved, articulated curriculum, eliminating population size and curriculum as variables. All correlations for Research Question 3 were significant at the .01 level.

Research Question 4

Determine the concurrent, external, construct validity of the ESPA using the CTP III as an external criterion measure

The fourth research question was to determine if the ESPA, first administered as a fourth-grade test in 1997 with cut-off scores set in 1999, was valid as determined through a correlation to a recognized, independently validated, standardized test taken by the same students within a month in 1999 and measuring the same content areas. Nine null hypotheses were used to test the concurrent, external aspect of construct validity of the ESPA as determined through a Pearson r between

the fourth-grade test, administered in May 1999, and the CTP III, Level D test, administered in April 1999:

- H_{O22}: The correlation between the May 1999 ESPA Language Arts Literacy test and the April 1999 CTP III, Level D Verbal Ability test is between 0 and .59.
- H_{O23}: The correlation between the May 1999 ESPA Language Arts Literacy test and the April 1999 CTP III, Level D Reading Comprehension test is between 0 and .59.
- H_{O24}: The correlation between the May 1999 ESPA Reading subtest test and the April 1999 CTP III, Level D Verbal Ability test is between 0 and .59.
- H_{O25}: The correlation between the May 1999 ESPA Reading subtest and the April 1999 CTP III, Level D Reading Comprehension test is between 0 and .59.
- H_{O26}: The correlation between the May 1999 ESPA Writing subtest and the April 1999 CTP III, Level D Verbal Ability test is between 0 and .59.
- H_{O27}: The correlation between the May 1999 ESPA Writing subtest, poem, and the April 1999 CTP III, Level D Verbal Ability test is between 0 and .59.
- H_{O28}: The correlation between the May 1999 ESPA Writing subtest, picture, and the April 1999 CTP III, Level D Verbal Ability test is between 0 and .59.

- H_{o29}: The correlation between the May 1999 ESPA Mathematics test and the April 1999 CTP III, Level D Quantitative Ability test is between 0 and .59.
- H_{o30}: The correlation between the May 1999 ESPA Mathematics test and the April 1999 CTP III, Level D Mathematics test is between 0 and .59.

The fourth research question measured the validity of the May 1999 ESPA as determined through a correlation with the April 1999, CTP III, Level D. The two tests were administered within a one month period. Data in Table 5 were used to capture H_{o22} – H_{o30}.

Table 5

Fourth Grade Class of 1999

April 1999 CTP III, Level D / May 1999 ESPA Correlations

ESPA	CTP III	N	Pearson r	r^2
LA Literacy	Verbal Ability	253	0.58	.34
LA Literacy	Reading Comprehension	253	0.62	.38
Reading	Verbal Ability	253	0.58	.34
Reading	Reading Comprehension	253	0.60	.36
Writing	Verbal Ability	253	0.35	.12
Writing(poem)	Verbal Ability	253	0.26	.07
Writing (picture)	Verbal Ability	253	0.31	.10
Mathematics	Quantitative Ability	251	0.77	.59
Mathematics	Mathematics	251	0.81	.66

The correlations in Table 5 ranged from a low of .26 to a high of .81. Two null hypotheses were rejected, both impacting the Mathematics test, thereby implying that validity was established for one ESPA test based on CTP III criteria. With similar correlations, null hypotheses H_{o30}, which stated that the correlation between the May 1999 ESPA Mathematics test and the April 1999 CTP III, Level D, Mathematics test was between 0 and .59, and H_{o29}, which stated that the correlation between the May 1999 ESPA Mathematics test and the April 1999 CTP III, Level D, Quantitative Ability test was between 0 and .59, were both rejected. The correlation between the two tests in H_{o30} was .81; the correlation between the two tests in H_{o29} was .77. Both correlations identified extremely strong and positive relationships. This replicated the relationships found between similar tests in Table 4. The r^2 for H_{o30}, a direct measure of the strength of the relationship, was .66. This identified that the relationship was two-thirds that of a perfect correlation. The r^2 for H_{o29} was .59, indicating that 59% of the variance could be explained by the two tests. Based on the data in Table 5, the ESPA Mathematics test was valid as demonstrated through the concurrent, external aspect of construct validity. Therefore, the ESPA Mathematics test was an appropriate measure of the discipline.

The null hypotheses were supported for H_{O22}, H_{O24}, and H_{O26} - H_{O28} with correlations ranging from .26 to .58. Two correlations were .58. Both H_{O22}, which stated that the correlation between the May 1999 ESPA Language Arts Literacy test and the April 1999 CTP III, Level D, Verbal Ability test was between 0 and .59 and H_{O24}, which stated that the correlation between the May 1999 ESPA Reading subtest test and the April 1999 CTP III, Level D, Verbal Ability test was between 0 and .59, were supported with a Pearson r of .58. While the correlations were positive, they were not strong enough to demonstrate validity. The r^2 was for each was .34, indicating that the strength of the relationship in each case was approximately one-third that of a perfect correlation with 34% of explained variability. At the same time, two-thirds of the variance could be attributed to chance, error, or external factors. It should be noted that both correlations involved the same CTP III test, Verbal Ability, and that the ESPA Reading subtest, H_{O22}, was a component of the ESPA Language Arts Literacy Test, H_{O22}.

The three Writing correlations were extremely low, ranging from .26 to .31. H_{O26}, which stated that the correlation between the May 1999 ESPA Writing total subtest and the April 1999 CTP III, Level D, Verbal Ability test was between 0 and .59, was supported with a very low correlation of .35. The r^2 of .12

indicated a relationship approximately one-third as strong as that found in H_{O22} and H_{O24} and less than one-fifth as strong as that found in H_{O29} and H_{O30}. H_{O28}, which stated that the correlation between the May 1999 ESPA Writing subtest, picture, and the April 1999 CTP III, Level D, Verbal Ability test was between 0 and .59, was supported with a Pearson r of .31. Again, while positive, this was very low. The r^2, which provides a direct measure of the strength of a relationship, was a weak .10. Finally, H_{O27}, which stated that the correlation between the May 1999 ESPA Writing subtest, poem, and the April 1999 CTP III, Level D Verbal Ability test was between 0 and .59, was supported with a low correlation of .26. With an r^2 of .07, it was apparent that the relationship was not strong. In fact, 93% of the variance could be attributed to chance, error, or external factors.

Two null hypotheses were referred for further study. H_{O23}, which stated that the correlation between the May 1999 ESPA Language Arts Literacy test and the April 1999 CTP III, Level D, Reading Comprehension test was between 0 and .59, found a correlation of .62. At the same time, H_{O25}, which stated that the correlation between the May 1999 ESPA Reading subtest and the April 1999 CTP III, Level D, Reading Comprehension test was between 0 and .59, found a correlation of .60. While both were

below the acceptable Pearson r for a validity measure, .70, each fell into the range for further review. However, since both appeared in the bottom range of the .60 category, and since the second correlation on both tests (H_{o22} and H_{o24}) was .58, the possibility of either test being valid is very questionable. At the same time, the review should consider the fact that the ESPA Reading correlations, .58 and .60, were higher than the GEPA Reading correlations.

As with the GEPA, the study raised a question as to the appropriateness of the use of ESPA as an accountability indicator for students and districts in Language Arts Literacy, especially writing. Again, it should be noted that the population was large for a correlation study, the students in the data population were housed in one building, and the fourth-grade had a consistent, written, board-approved, articulated curriculum, eliminating population size, location, and curriculum as variables. All correlations for Research Question 4 were significant at the .01 level.

Summary of the Data

The validity study of the HSPT11 indicated that the HSPT11 Reading and Mathematics tests demonstrated concurrent, external, construct validity and were, therefore, appropriate indicators

of high-stakes student decisions. Approximately half the variance was attributable to the two tests in the correlation. However, the HSPT11 Writing test and Writing Task (Essay) component did not demonstrate concurrent, external, construct validity. For the Writing test, approximately three-quarters of the variance in the validity correlation was due to chance, errors, or other factors while, for the Essay, this jumped to ninety percent of the variance. The data in Table 1 and Table 2 indicate that neither the Writing test nor the Writing Task should be used for high-stakes decisions.

The Mathematics EWT demonstrated strong predictive validity with HSPT11, indicating that student decisions based on the scores were appropriate. The Writing test demonstrated poor validity, raising questions about score-based student and curriculum decisions in this content area. As with the concurrent, external aspect of the construct validity data listed in Tables 1 and 2, the variance due to chance, error, or external factors in the latter predictive study was extremely high, especially for the Writing Task (Essay). With a predictive validity correlation of .67, the Reading test needs further review. However, it should be noted that over half the variance was due to chance, error or other factors, raising questions about the educational value of decisions based on this test.

Both the ESPA and the GEPA Mathematics tests demonstrated strong, concurrent, external, construct validity, indicating that they were appropriate tests on which to base student decisions. Correlations on the GEPA Reading and Writing tests indicated poor concurrent, external, construct validity. This finding was supported by the high percentage of variance due to chance, error, or external factors, raising concerns about their use for student placement and curriculum decisions.

The correlations for the ESPA Reading test indicated the need for further study. At the same time, the extremely low correlations on the ESPA Writing tests identified a specific area for caution, one that should draw immediate attention from test developers.

Chapter V

Summary, Conclusions, and Recommendations

Introduction

This chapter is divided into three sections. The first section presents a summary of the study, the second section discusses the results and conclusions, and the final section addresses recommendations for future state assessments and future validity studies.

Summary

The purpose of this study was to determine the concurrent, external aspect of construct validity for three tests in the New Jersey state assessment program - Elementary School Proficiency Assessment (ESPA), Grade Eight Proficiency Assessment (GEPA) and High School Proficiency Test (HSPT11) by correlating scores for the same students on the state tests and on grade and content appropriate, national, standardized tests. In addition, the study looked at the predictive validity of the EWT by correlating the scores for the same students on both the EWT and the HSPT11, the latter taken three years later in junior year of high school.

Results

Research Question 1

A. Determine the concurrent, external, construct validity of the HSPT11 using the PSAT as an external criterion measure

B. Determine the concurrent, external, construct validity of the HSPT11 using the SAT as an external criterion measure

Based on the data presented in Chapter IV, concurrent, external validity was found for the HSPT11 Mathematics test. The correlation between the HSPT11 Mathematics test and the PSAT Mathematics test, as well as that between the HSPT11 Mathematics test and the SAT Mathematics test, was .72. This exceeded the validity standard established as r=.7.

Discussing validity, Gronlund (1981) stated, "The scores of any particular test can be expected to correlate substantially with the scores of other tests that presumably measure the same thing...For any given test, we would predict higher correlations with like tests and lower correlations with unlike tests" (Gronlund, 1981, p.83). The correlations between the mathematics tests support this premise. As stated by Cronbach (1971) and Cizek (1998), the strength of this correlation further indicated that the conclusions were meaningful, accurate, and useful. Based on the data, it can be stated that the HSPT11 Mathematics

test was a valid measure of the mathematical concepts that should be known at the completion of thirteen years of schooling. As such, the HSPT11 Mathematics test is an appropriate assessment instrument for high-stakes decisions that have accountability for students and districts.

The correlations for the HSPT11 Reading test followed a similar pattern. The Pearson r between the HSPT11 Reading test and the PSAT Verbal test was .71; the Pearson r between the HSPT11 Reading test and the SAT Verbal test was .77. Each identified a strong, positive relationship between the state assessment and the national, standardized test. Based on the data described in Chapter IV, the correlations for the HSPT11 Reading test exceeded the validity standard established as r=.7. The data provided by the two correlations indicated that the HSPT11 Reading test was a valid assessment instrument. From the Pearson r, it can also be stated that the HSPT11 Reading test was an effective measure of developed reading skills and an appropriate assessment for high-stakes decisions that have accountability for students and districts.

Recognizing that the PSAT and SAT are connected, with the same managers for the verbal and mathematics components, the parallel performance was expected. However, since the PSAT was administered within two weeks of HSPT11, while the SAT was

administered at a point within one year of HSPT11, higher correlations would be expected between HSPT11 and PSAT. Therefore, the unexpected, slightly higher correlations between the HSPT11 Reading test and its SAT counterpart must be noted.

The correlations for the HSPT11 Writing test presented a different picture. The correlation between the HSPT11 Writing test and the PSAT Verbal test was .51; that between the HSPT11 Writing test and the SAT Verbal test was .50; and that between the HSPT11 Writing Test and the PSAT Writing Test was .49. The closeness of the numbers must be noted. In each case, the correlation was positive. However, none of the three could be interpreted as strong. Based on the validity standard established as r=.7, the HSPT11 Writing test was not a valid assessment instrument. The data in this study were similar to those reported by Herman and Winters (1994) during a review of writing portfolios. According to the researchers, "One useful approach in determining what portfolio scores mean is to look for patterns of relationships between the results of portfolio assessments and other indicators of student performance. Score meaning becomes supported when portfolio scores relate highly to other, valued measures of the same capability. Using this approach...the researchers found moderate correlations ranging

from .47 to .58 between writing portfolio scores and direct writing assessments" (p. 51).

Specific attention needs to be paid to the HSPT11 Writing Task (Essay). The writing correlation between the HSPT11 Essay test and the 1997 PSAT Writing test was a low, weak .33. This number was far below the validity standard of r=.7. In fact, the correlation was so low that the abilities tested would have to be independent. The findings in this study were similar to those described by Herman and Winters (1994), "Gearhart and others (1993) found virtually no relationship when comparing results from writing portfolios with those from standard writing assessments. In fact, two-thirds of the students who would have been classified as "masters" based on the portfolio assessment score would not have been so classified on the basis of the standard assessment" (p. 51).

Based on the data from the four correlations, neither the total HSPT11 Writing test nor the Writing Task (Essay) component was a valid assessment instrument. The very low correlation and coefficient of determination for the Writing Task (Essay) raised particular concerns about the appropriateness of this test as an assessment instrument with accountability for districts and students. In a Sunday Star-Ledger article, former New Jersey education commissioner Klagholz acknowledged "Writing is hard to

assess. I wouldn't say that they are bad tests, but the tests are evolving and the evolution isn't as good as in reading and math" (Alaya, 1999, p. 29).

Research Question 2

Determine the predictive validity test of the EWT using the HSPT11 as the criterion measure

The second research question was to determine if the EWT did predict performance (predictive validity) on the HSPT11, laying a foundation for questioning if the GEPA can predict performance on the HSPA and, as an extension, if the fourth-grade ESPA can predict performance on the eighth-grade GEPA. According to Messick (1993), the construct validity framework allows a rational basis for prediction. "It leads us to address, as well, varieties of discriminant evidence essential in the construct validation of both predictor and criterion measures (p. 77). Quoted by Messick, Guion (1976) stated that the predictive hypothesis was "the outcome of a rational process linking the domain theory to the choice of criterion and predictor constructs as well as to the empirically grounded construct interpretations of the criterion and predictor measures." He continues that what is appraised in predictive validity is the "validity of the hypothesis of a relationship between the test and a criterion measure" (p. 77). In perhaps

more direct words, Cronbach (1984) described the need to compare a test to the prediction for a "straightforward empirical check on the value of the test for predictive validity" (p.103). According to Anastasi (1982), predictive validity is required when there is interest in predicting or determining the relationship between two measures over an extended period of time. "If we want to use test scores to predict outcome in some future situation, such as an applicant's performance in college, we must use tests with high predictive validity against the specific criterion" (p. 30). As with concurrent validity, the two sets of data must always be on the same individual.

Predictive validity answers the challenge of the EWT, and potentially the GEPA and ESPA. The skills identified for the EWT were the benchmarks for those on HSPT11. In Department of Education meetings throughout New Jersey, it was stated that the test should be used to identify students who might have difficulty passing the high-stakes HSPT11. As with the EWT and HSPT11, the same relationship is expected between the GEPA and the, in development, HSPA as well as between the GEPA and ESPA. According to Gronlund (1981), "If the results are to be used to predict student success in some future activity, we should like them to provide as accurate an estimate of future success as possible" (p. 65).

To determine predictive validity, the March 1995 EWT was correlated to the performance of the same students on the October 1997 HSPT11. Four specific research questions supported the second major research area. The specific questions, tested as null hypotheses, allowed the division of the major research area into correlations between aligned tests.

The correlation between the EWT Mathematics test and the HSPT11 Mathematics test was .79, indicating a very strong, positive relationship. This was appropriate since the Eighth-Grade Mathematics Skills Committee patterned the EWT after the HSPT11. Therefore, repeating a previous quote from Gronlund (1981), "The scores of any particular test can be expected to correlate substantially with the scores of other tests that presumably measure the same thing" (p.83). From the correlation it can be stated that the two mathematics tests "presumably measure the same thing". Based on the Pearson r, the EWT Mathematics test exceeded the validity standard of r=.7, demonstrating predictive validity with the HSPT11 Mathematics test. According to the data, the EWT Mathematics test predicted performance on the HSPT11 Mathematics test. From the strong correlation, it was further apparent that the EWT Mathematics test was an appropriate indicator for student and district accountability. This conclusion was strengthened through a

regression analysis that indicated that a 1997 EWT Mathematics passing score of 100 was equivalent to a 1999 HSPT11 Mathematics passing score of 336. The regression analysis supported the observation that the EWT Mathematics test served as a good indicator for predicting success on the HSPT11.

The correlation between the 1995 EWT Reading test and the 1997 HSPT11 Reading test fell just outside the range of predictive validity (.67) but close enough to warrant further review according to the descriptors of this study. However, a regression analysis supported the probability that the 1997 EWT Reading test was not a good predictor of a passing or failing score on the 1999 HSPT11 Reading. According to the regression analysis on the Class of 1999 population, a minimal EWT Reading passing score of 100 led to a failing score of 171.79 on the 1999 HSPT11 Reading. Therefore, even with moderate, predictive validity, the EWT Reading test did not appear to be an effective predictor for passing or failing the similar test of the HSPT11 and should be considered questionable as an indicator for student and district accountability.

The correlation between the 1995 EWT Writing test and the 1997 HSPT11 Writing test, .47, fell well below the acceptable point for predictive validity. It should be noted that the correlation coefficient was similar to those found in the three

concurrent, external, construct validity tests (.49-.51) on the HSPT11. While positive, this number did not represent a strong relationship.

The questions raised on the total EWT Writing Test were intensified by the data identified through the correlation between the 1995 EWT Writing Task (Essay) and the 1997 HSPT11 Writing Task (Essay). This Pearson r was a very low .22. The correlation was so low that the abilities tested would have to be independent. Based on the validity standard of r=.7, it could be stated that the EWT Writing Task (Essay) did not demonstrate predictive validity and was not an appropriate accountability indicator for student performance on the HSPT11 Writing Task (Essay).

As with the writing test as a whole, the predictive, external correlation for the essay was similar to the data listed for Research Question 1. The correlation between the HSPT11 Writing Task (Essay) and the PSAT Writing Test was .33; the correlation between the EWT Writing Task (Essay) and the HSPT11 Writing Task (Essay) was .22. The consistency of the two, extremely low correlations strengthened the fact that the New Jersey essay tests did not demonstrate validity and need a close review before continued implementation or replication.

A regression analysis on the writing tests led to another interesting dimension. According to the regression analysis on the Class of 1999 population, a minimal EWT Writing passing score of 100 led to a passing score of 336 on HSPT11, similar to the regression for the Mathematics test. This result further demonstrated the instability of the Writing test. Writing is the last communication skill to develop. Students should be able to read before they write. Therefore, the reading scores should be stronger than the writing scores. This is consistent with the predictive validity correlations. However, the regression analysis leads to further questions about the use of the state writing assessment as an indicator of future performance.

The data for both the total Writing test and the Writing Task (Essay) component were inconsistent with Department of Education statements. The EWT was developed as a result of the 1988 HSPT law to identify students in need of remedial education services and to determine the effectiveness of the elementary curriculum in preparing students for the skills assessed by HSPT11. As with HSPT11, the New Jersey Department of Education described the EWT as "a rigorous test of essential skills in reading, mathematics and writing" (New Jersey Department Of Education, 1997b, p. 3). According to a number of documents published by the New Jersey Department of Education, students

had to first master the EWT skills before mastering those on HSPT11 (New Jersey Department of Education, 1989, 1990a, 1990b, 1997c). Furthermore, the two tests had the same core of item developers and a consistent format. The EWT committee was composed of educators who developed the eleventh grade skills, along with additional elementary school representatives. The final product for both tests included multiple-choice and free-response items, the latter requiring students to construct written responses. Consistent skills were tested with those for the EWT at a lower level of complexity and sophistication. Therefore, there was a parallel design for the two tests, each appropriate to the respective grade levels. Based on this information, a high correlation between the two tests was expected. The data did not support the expectation. From the data, it was concluded that the EWT Writing test was not a good predictor of performance on the HSPT11 Writing test. Therefore, the EWT Writing test was not an appropriate indicator for student and district accountability.

Of particular note was the extremely low correlation between the two essay tests. Both used the same prompt format; both called for a persuasive essay. In addition, HSPT11 and EWT were scored with the same rubric under the auspices of the same scoring company. According to Peter Peretzman, spokesman for

former commissioner Leo Klagholz, the state had not changed the scoring rubrics over the years (Alaya, 1999). Attention to the low predictive validity is important since the New Jersey Department of Education has announced the same predictive role for the new ESPA and GEPA tests, again incurring accountability for students and districts.

Research on essay tests indicated that scoring may be a major variable contributing to the unexpectedly low correlations. Concern about scoring open-ended or performance items has been voiced by many experts. This concern is not new. In 1889, F. Y. Edgeworth was the first to research essay test score validity. More recently, Herman & Winters (1994) stated, "Raters who judge student performance must agree regarding what scores should be assigned to students' work within the limits of what experts call 'measurement error...' Do they (raters) assign the same or nearly similar scores to a particular student's work? If the answer is no, then student scores are a measure of who does the scoring rather than the quality of the work" (p. 49). LeMahieu (1995b) echoed this thought, noting concerns about the probability of obtaining acceptable levels of agreement between judges and whether the rubrics support sufficiently high expectations for students. According to LeMahieu, instability of

Judging student work on factors irrelevant to the quality of the performance was also a concern. Eha (1998) voiced questions about "the numerous factors-such as differential application of scoring rubrics and assorted 'halo effects'-that add additional error to estimates of student ability" (p.1). According to Wiggins (1994), scorers tend to over-emphasize process and form criteria in scoring performance and under-emphasize or ignore impact criteria-the criteria that relate to purpose and desired effects.

Substantiating the theorists, in a review of the Vermont state assessment program, researchers Daniel Koretz, Brian Stecher, Stephen Klein and Daniel McCaffrey of the RAND Corporation found that it was hard to train large numbers of raters at a sufficient level of accuracy (Bracey, 1995). Vermont was not alone with this problem. Major publishers, such as Harcourt, had similar experiences. In 1998, the Texas publisher had to rescore the fourth and eighth grade writing sections for both the Vermont and Rhode Island assessments.

The New Jersey open-ended writing assessments were scored by Measurement Inc of Durham, North Carolina. According to an article in The Record (Glovin, 1998), Measurement Inc. hired college-educated jobbers for $7.25-$7.75 per hour to score the essays. Up to seventy temporary workers, ranging from a former

fighter pilot to an artist launching a gallery, graded an average of 150 papers in a seven-hour day. A $200.00 bonus was paid after 8,000 papers. Readers did spend three days learning the grading scale designed by New Jersey. However, in a telling interview, one four-year reader, Julian Harrison, who felt pressured to earn the $200.00 bonus, remembers the following:

> There were times I'd be reading a paper every 10 seconds. It was horrific...you could actually-I know this sounds very bizarre-but you could put a number on these things without actually reading the paper... (wrong grades) Either I read it too fast or I didn't recognize what the child [meant] or maybe I got impatient because the child's handwriting was very bad (p.8).

Based on the low correlations found in this validity study, the New Jersey state assessment program needs to review its scoring practices. As noted by Alan E. Farstrup, executive director of the International Reading Association, "This raises a big issue of how quickly and how superficially should we rely on performance-based, high-stakes data when we simply don't know if it's telling us what we need. Some very serious decisions are being made about kids' lives based on the

instruments we've pressed into service..The consequences could be devastating" (Manzo, 2000, p. 17).

Research Question 3

Determine the concurrent, external, construct validity of the GEPA using the CTP III as an external criterion measure

The third research question was to determine if the eighth-grade GEPA, first administered in March 1999, was valid as determined through correlation to a recognized, independently validated, national, standardized test, taken by the same students within a one-year time period and measuring the same content areas. Nine specific research questions supported the third research topic. The nine specific questions, tested as null hypotheses, studied the concurrent, external aspect of construct validity of the GEPA as determined through a Pearson r between the eighth grade test, administered in March 1999, and the CTP III, Level E test, administered in April 1998. Testing the specific questions as null hypotheses allowed the division of the major research question into correlations between aligned tests.

The correlation of the GEPA Mathematics Test with the CTP III Quantitative Ability Test and with the CTP III Mathematics Test provided similar, strong, positive results. The Pearson r

between the GEPA Mathematics Test and the CTP III Quantitative Ability Test was .83; the Pearson r between the GEPA Mathematics Test and the CTP III Mathematics Test was .82. This result was similar to the .86 found by Eha (2000) in a correlation of CTP III, Level E Mathematics and the New York state eighth-grade mathematics assessment. As stated by Eha, the correlations were what would be expected in a reliability study when two different, but parallel, forms of the same test are correlated. This reinforced the fact that the tests measured the concept of mathematics appropriate to the grade level. Cronbach (1971) and Cizek (1998) noted that a strong correlation also indicated that the conclusions yielded by the data were meaningful, accurate and useful. Based on the validity standard established as r=.7, the GEPA Mathematics Test demonstrated concurrent, external construct validity and was a good measure of the discipline. It was also concluded that the test was an appropriate indicator for student and district accountability. It should be noted the CTP III tests were completely multiple-choice while the GEPA had a variety of response formats. From the high correlations, it could be assumed that response format did not impact on validity. This is consistent with the work of Gearhart et al (1993) and that of Koretz et al (1993, 1994).

The results of the GEPA Language Arts Literacy test and its Reading and Writing components painted a different picture. The GEPA Language Arts Literacy Test was correlated to the following CTP III tests: Verbal Ability, Reading Comprehension, and Writing Process. The Reading component of the GEPA Language Arts Literacy test was correlated to both the CTP III Verbal Ability Test and the Reading Comprehension Test while the Writing component was correlated to both the CTP III Verbal Ability Test and the Writing Process Test. The correlations ranged from .41 to .59; indicating positive, but not strong, relationships. Repeating Gronlund (1981), "The scores of any particular test can be expected to correlate substantially with the scores of other tests that presumably measure the same thing...For any given test, we would predict higher correlations with like tests and lower correlations with unlike tests" (p.83). Based on this statement, it was apparent that the GEPA Language Arts Literacy, Reading, and Writing tests did not measure the concepts as tested in the similar level CTP III tests. Since the CTP III is an established and recognized test of Verbal Ability, Reading Comprehension and Writing Process, the low correlations indicated that the GEPA did not measure verbal ability skills, reading comprehension, or knowledge of writing process.

The data for the GEPA test was inconsistent with the previously stated data for the HSPT11 and the EWT. However, the skills measured on the GEPA were not aligned to the skills on the two former tests. Instead, GEPA is a measure of the CCCS.

Is that the problem? Analyses of standards are limited and often self-serving. The American Federation of Teachers (AFT), the Council for Basic Education (CBE), and the Thomas B. Fordham Foundation have each reviewed standards. Of these, the Fordham Foundation, based in Washington and headed by Chester E. Finn Jr., a former assistant US secretary of education under President Reagan, provided the best example of a consistent, structured and nationally recognized evaluation of state standards in five core academic areas. For Language Arts Literacy, the New Jersey CCCS scored an F. At the same time, the Mathematics Standards rated a C (Finn et al, 2000). Looking back to an earlier, July 1997, Fordham report on State English Standards, Dr. Sandra Stotsky, research associate at both the Harvard Graduate School of Education and the Boston University School of Education, wrote that the New Jersey Language Arts standards had "many limitations...Its standards for reading, literary study, and writing are weak...Many standards lack specificity and measurability..." Her recommendation was that "The document needs to be completely rewritten...with specific

and measurable standards " (p. 61). This supported the 1996 American Federation of Teachers analysis which stated that New Jersey failed to meet its common core criterion and gave passing grades only to the state's mathematics and science standards (Mooney, 2000b).

Standards do not create the only question with GEPA. Research suggests that a problematic factor with the GEPA Language Arts Literacy test could be due to the method of determining the scores. According to the New Jersey Department of Education (1999d), proficiency levels were determined by panelists who represented the various DFGs and regions of the state. As with ESPA, the panelists were either practicing teachers or curriculum supervisors in one of the content areas. A holistic classification method was used for the proficiency-level setting study. Based on thirty-three student test booklets, covering the range of student performance as determined by the testing company, the panelists individually classified each booklet as partially proficient, proficient or advanced proficient. This process was similar, but broader, than that used in setting NAEP achievement levels. Naep, HSPT11, and EWT scores were determined through a modified Angoff approach. Angoff is a method of scoring where knowledgeable judges rate each item on the test, estimating the proportion of marginal

examinees who would correctly answer the question. Under the GEPA format, judges looked holistically at the entire booklet, rather than each item. In practice, the holistic procedure employs greater subjectivity than an item approach. A study released in the fall of 1999 by the National Research Council described the NAEP modified Angoff process as fundamentally flawed (Hoff, 1998c, h; Manzo, 1999b). Based on the fact that an item review was considered flawed, the conclusion would be that a more holistic approach would offer even greater flaws.

The holistic approach was also implemented for scoring the Mathematics test. Why would it be a factor in Language Arts Literacy and not Mathematics? The answer lies back in the standards. In Mathematics, recognition of an Advanced Proficient, Proficient or Partially Proficient response is based on measureable, defined standards. Identification of categories is more difficult with the broad Language Arts Literacy standards which lack quantifiable indicators.

In addition, the Language Arts Literacy test subsumes the writing component. As noted in the discussion of Reasearch Question 2, writing is difficult to score. Concern about essay test score validity has been voiced since the work of F. Y. Edgeworth in 1889. More recently, at its March 2000 meeting, the NAGB voted to exclude the results of the 1999 NAEP writing

assessment from the trend report scheduled to be released in summer 2000. Results from the 1994 and 1998 writing tests will also be removed from the NAEP website. Gary W. Phillips, acting commissioner of the National Center for Education Statistics, recommended the move after learning that there were errors associated with the scaling model used for scoring. In general, concern has been raised about the reliability of this format in statewide and national testing (Manzo, 2000). Gary Phillips, acting commissioner of the National Center for Education Statistics, stated "I've just lost confidence that the data are reliable" (Manzo, 2000, p. 1). According to experts, assessing writing with only a few questions is problematic. In the same Manzo article, Stephen Klein of the RAND corporation said, "Students do better on some prompts than others, so you need a large number of prompts to get a clear picture of how a student is doing" (p. 17).

Based on the data, the GEPA Language Arts Literacy test, and the Reading and Writing components, did not meet the validity standard of r=.7, nor the review range of r=.6-.69. It was, therefore, concluded that they were not valid instruments. This raised a question as to the appropriateness of their use, especially since the GEPA has accountability issues for students and districts. Placing students in a basic skills Language Arts

program as a result of GEPA scores needs to be reviewed. Of particular note was the fact that the GEPA Reading test had low correlations of .47 and .53, indicating an area of concern not raised with the EWT.

It should be stated that the population was large for a correlation study and that the students in the data population had a consistent, written, board-approved, articulated curriculum, eliminating population size and curriculum as variable.

Research Question 4

Determine the concurrent, external, construct validity of the ESPA using the CTP III as an external criterion measure

The fourth research question was to determine if the ESPA, first administered as a fourth-grade test in 1997 with cut-off scores set in 1999, was valid as determined through a correlation to a recognized, independently validated, standardized test taken by the same students within a month in 1999 and measuring the same content areas. Nine specific research questions supported the fourth research area. The nine specific questions, tested as null hypotheses, studied the concurrent, external aspect of construct validity of the GEPA as determined through a Pearson r between the fourth grade test,

administered in May 1999, and the CTP III, Level D test, administered in April 1999. Testing the specific questions as null hypotheses allowed the division of the major research question into correlations between aligned tests.

The correlation of the ESPA Mathematics Test with the CTP III Quantitative Ability Test and with the CTP III Mathematics Test provided similar, strong, positive results. The Pearson r between the ESPA Mathematics Test and the CTP III Quantitative Ability Test was .77; the Pearson r between the ESPA Mathematics Test and the CTP III Mathematics Test was .81. It was concluded from the two correlations that the tests measured the same concept of mathematics. Based on the validity standard established as r=.7, the ESPA Mathematics test was a valid assessment instrument.

As with the correlations for the eighth-grade mathematics tests, the data produced numbers that would be expected in reliability studies when two different, but parallel, forms of the same test are correlated. This intensified the conclusion that the tests measured mathematics as appropriate for a fourth-grade population. Based on the data, the ESPA Mathematics Test demonstrated concurrent, external, construct validity and was a good measure of the discipline. It was also concluded that the test was an appropriate indicator for student and district

accountability. It should be noted the CTP III tests were completely multiple-choice while the ESPA had a variety of response formats. From the high correlations, it could be assumed that response format did not impact on validity. This study supported the results of the concurrent, external aspect of construct validity study for the eighth-grade GEPA.

The results of the ESPA Language Arts Literacy test and its Reading component indicated an area for further review. The ESPA Language Arts Literacy Test and its reading component were independently correlated to the following CTP III tests: Verbal Ability and Reading Comprehension. The correlations for the Language Arts Literacy test were .58 and .62 and those for the Reading component were .58 and .60. It should be noted that the correlations were positive and similar to the .61 found by Eha (2000) in the New York fourth-grade study.

While the standard for validity was established as r=.7, a review range was established as r=.6-.69. Therefore, one correlation for each test fell into the area for review, although on the low end, warranting further study. However, both correlations appeared in the bottom range of the .60 category with the second correlation on each test a .58. With these numbers, it can be projected that the possibility of either test being valid is very questionable. In a recent article, E. D.

Hirsch Jr. (2000) stated that "All of the well-established reading tests are valid, reliable, and highly correlated with one another" (p. 40). That was not found with this data. The CTP III is an established and recognized test of Verbal Ability and Reading Comprehension. Therefore, even with the .60 and .62 correlations, it can be stated that the ESPA does not adequately measure verbal ability skills or reading comprehension.

The major area for concern was found in the three correlations for writing. A Pearson r was derived for the ESPA Writing test as a total and for each of the components (poem and picture) through correlations with the CTP III Verbal Ability Test. The correlations ranged from .26 to .35, indicating positive, but very weak relationships. Based on validity standard of r=.7, neither the ESPA Writing test nor any subtest was valid. The data were consistent with the Writing Task correlations for the HSPT11 and the EWT.

Both the ESPA and GEPA were developed to measure the CCCS. As stated in the GEPA discussion, the American Federation of Teachers (AFT), the Council for Basic Education (CBE), and the Thomas B. Fordham Foundation each reviewed standards. Each found the Language Arts Literacy Standards unacceptable. Obviously, there is a problem with the domain being tested.

Based on the data, the ESPA Language Arts Literacy test, and Reading and Writing components, did not demonstrate concurrent, external, construct validity. This raised a question as to the appropriateness of their use, especially since the ESPA has accountability issues for students and districts. Placing students in a basic skills Language Arts program as a result of ESPA scores needs to be reviewed. In particular, the extremely low correlations (.26-.35) for the ESPA writing raised a strong area for concern.

As with the GEPA, the study raised a question as to the appropriateness of the use of ESPA as an accountability indicator for students and districts in Language Arts Literacy, especially writing. Again, it should be noted that the population was large for a correlation study, the students in the data population were housed in one building, and the fourth-grade had a consistent, written, board-approved, articulated curriculum, eliminating population size, location, and curriculum as variables.

Summary

All mathematics tests in the New Jersey state assessment program demonstrated external validity as determined through a correlation with a recognized, national, standardized test. The mathematics tests of HSPT11, GEPA, and ESPA demonstrated

concurrent, external, construct validity while the mathematics test of the EWT proved to have predictive validity. The conclusion on the latter was supported by a regression analysis. Therefore, it can be stated that the New Jersey state mathematics tests are good measures of the content area and appropriate indicators for accountability decisions for students and districts.

The HSPT11 Reading test demonstrated concurrent, external, construct validity. The test was, therefore, a good instrument for high-stakes student decisions. The EWT Reading test was referred for further study. The Pearson r fell just below the .7 required for validity. In addition, a regression analysis supported the probability that the 1997 EWT Reading test was not a good predictor of a passing or failing score on the 1999 HSPT11 Reading. Therefore, even with moderate, predictive validity, the EWT Reading test did not appear to be an effective predictor for passing or failing the similar test of the HSPT11. Neither the GEPA Language Arts Literacy test nor the Reading component demonstrated concurrent, external, construct validity, raising questions about their use for student and district decisions.

The ESPA Language Arts Literacy test and the Reading component were also referred for further study with one

correlation in each area in the review category of .60-.69. It should be noted that both correlations appeared in the bottom range of the ..60 category with the second correlation on each test a .58. With these numbers, it can be projected that the possibility of either test being valid is unlikely. The CTP III is an established and recognized test of Verbal Ability and Reading Comprehension. Therefore, even with the .60 and .62 correlations, it can be stated that the ESPA does not adequately measure verbal ability skills or reading comprehension.

That statement is strengthened through a review of the domain for the GEPA and ESPA tests. The HSPT11 and EWT Reading tests measured skills identified by aligned state Reading Skills Development Committees. The GEPA and ESPA measured reading as determined through the CCCS. A concern was raised about the latter. The American Federation of Teachers (AFT), the Council for Basic Education (CBE), and the Thomas B. Fordham Foundation have each reviewed standards. From each review, the Language Arts Literacy Standards were found to be unacceptable.

What happens when the domain is described by independent sources as a weak representation of the content area? Does a high correlation between the assessment and the standards indicate anything more than the fact that the inferior domain and the assessment measure the same thing? Would a review of the

two ESPA Reading correlations in the low .6 range change the fact that verbal ability and reading comprehension skills are not appropriately measured through the fourth-grade state assessment? Before conducting a review of the correlations for the two questions, the state should consider the independent ratings of the standards and review the domain for the ESPA and GEPA Language Arts Literacy and Reading assessments.

No writing test in the New Jersey state assessment program demonstrated external validity as determined through a correlation with a recognized, national, standardized test. Therefore, it can be stated that the New Jersey state writing tests are poor measures of the discipline and not appropriate indicators for accountability decisions for students and districts. For the ESPA and GEPA writing assessments, the criticism of the New Jersey Language Arts Literacy Standards, noted above, is a pertinent factor requiring attention from the state.

The low, open-ended Writing Task (Essay) correlations were of particular concern. For the HSPT11 and the EWT, the extremely low predictive, external correlation of .22 was unexpected. The HSPT11 and EWT had the same core of item developers and a consistent format with a writing "prompt". In addition, the two essay tests were scored with the same rubric under the auspices

of the same scoring company. While the ESPA and GEPA open-ended writing components had different formats, they, too, were scored with the same rubric under the auspices of the same scoring company. That may be the problem. As noted above, The Record (Glovin, 1998) uncovered the fact that New Jersey writing assessments were scored in North Carolina by college-educated, hourly workers who graded an average of 150 papers in a seven-hour day. Based on the low correlations found in this validity study, the New Jersey state assessment program needs to review not just the Language Arts Literacy Standards but also its scoring practices.

The review of ESPA and GEPA scoring should not be limited to writing assessments. It was noted previously that only seventy-seven tests were used to set the benchmarks in each test. The New Jersey Department of Education relied on a holistic approach. This process was similar, but broader, than that used in setting NAEP achievement levels. Naep, HSPT11, and EWT scores were determined through a modified Angoff approach. Angoff is a method of scoring where knowledgeable judges rate each item on the test, estimating the proportion of marginal examinees who would correctly answer the question. Under the GEPA and ESPA format, judges looked holistically at the entire booklet, rather than each item. In practice, the holistic procedure employs

than each item. In practice, the holistic procedure employs greater subjectivity than an item approach.

It should be noted that the population for each correlation study was large; for each grade-level, the students in the data population were housed in one building; and the district has a consistent, written, board-approved, articulated curriculum. This eliminated population size, location, and curriculum as variables. All correlations for the four research questions were significant at the .01 level.

Conclusions

- The HSPT11 Mathematics test demonstrated concurrent, external, construct validity and was, therefore, an appropriate indicator for high-stakes student decisions.
- The EWT Mathematics Test demonstrated predictive validity and was an appropriate predictor of performance on the HSPT11 Mathematics test.
- Both the GEPA Mathematics test and the ESPA Mathematics tests demonstrated concurrent, external, construct validity, indicating that they were appropriate tests for accountability decisions for students and districts.

- The HSPT11 Reading Test demonstrated concurrent, external, construct validity and was, therefore, an appropriate indicator for high-stakes student decisions.
- Neither the GEPA Language Arts Literacy Test nor the GEPA Reading Test demonstrated concurrent, external, construct validity, indicating that they were not appropriate tests on which to base accountability decisions for students and districts.
- Serious questions were raised about the New Jersey Language Arts Literacy Standards that served as the base for the ESPA and GEPA Language Arts Literacy Tests and its components.
- Strong concerns were raised about the holistic scoring format of the ESPA and GEPA tests and the inherent subjectivity.
- Serious concerns were raised about all writing tests and, in particular, the essay components. Neither the HSPT11 Writing Test, the GEPA Writing Test, nor the ESPA Writing Test demonstrated concurrent, external, construct validity, indicating that they were not appropriate tests on which to base accountability decisions for students and districts. The EWT Writing Test did not demonstrate predictive validity, indicating that it was not an appropriate predictor of performance on the HSPT11 Writing Test and, therefore, not an appropriate assessment for student placement decisions. In

particular, the open-ended writing components on each test had extremely low correlations, leading to a question about their continued use. Since the purpose of good assessment is to inform instruction and, at the same time, to provide students, parents, administrators, and the public with accurate and meaningful information about students' progress, it is important for the information to be accurate and meaningful. Based on the low correlations found in this validity study, the information provided to districts by the New Jersey Department of Education on writing performance lacked meaning. To correct this situation, the New Jersey state assessment program needs to review not just the Language Arts Literacy Standards but also its scoring practices.

- The EWT Reading test, along with the ESPA Language Arts Literacy Test and the ESPA Reading test, were referred for further study.

Recommendations for Future New Jersey State Assessments

The findings and conclusions of this study generate the following recommendations:

- The HSPT11, GEPA and ESPA Mathematics tests are all valid tests. The development process, format, and scoring processes

should be continued and replicated in the state assessment program.

- According to the data, the GEPA and ESPA Reading tests do not demonstrate external validity and are, therefore, not valid instruments. This information, coupled with independent analyses of the standards by the Fordham Foundation, AFT, and CBE, indicated that the standards must be reviewed for appropriateness and that the tests should be redesigned before being recognized as acceptable indicators for student and district accountability.
- The holistic scoring procedure for the ESPA and GEPA should be reviewed. This process was similar, but broader, than that used in setting NAEP achievement levels. Naep, HSPT11, and EWT scores were determined through a modified Angoff approach. A congressional report released in September 1998 called the NAEP process for measuring student achievement to be fundamentally flawed, the scoring subjective and not reflective of the results of similar large-scale tests such as the Advanced Placement Test. A more subjective approach would intensify these questions.
- According to the data, the EWT Reading Test did not demonstrate predictive validity. Although the test was referred for further study, the regression analysis supported

the concept that the test was not a good indicator of performance on the HSPT11 Reading test. Since this pairing is the prototype for the prediction of performance from the ESPA to the GEPA, and from the GEPA to the in-development HSPA, the concept needs to be further studied and refined before assigning student and district accountability consequences.

- Serious concerns were raised about all writing tests and, in particular, the essay components. Neither the HSPT11 Writing Test, the GEPA Writing Test, nor the ESPA Writing Test demonstrated concurrent, external, construct validity, indicating that they were not appropriate tests on which to base accountability decisions for students and districts. The EWT Writing Test did not demonstrate predictive validity, indicating that it was not an appropriate predictor of performance on the HSPT11 Writing test and, therefore, not an appropriate test for student placement decisions The open-ended writing components on each test had very low correlations, creating a question about their continued use. Based on the low correlations found in this validity study, the information provided to districts by the New Jersey Department of Education on writing performance lacked meaning. In addition, this raised questions about the predictability of ESPA and GEPA writing assessments. To correct this situation,

the New Jersey state assessment program needs to immediately review not just the Language Arts Literacy Standards but also its scoring practices.

- Finally, ensure the predictive validity between two measures before implementing predictive accountability.

Recommendations for Future Validity Studies

- The study included one DFG "I" district. Additional studies that include districts in DFGs "A" - "H," "J," and "V" would add further data to enhance the conclusions.
- The population of the study included only regular education students. With the emphasis on aggregate district scores, additional studies that include a special education population, or that focus on a special education population, could provide data to enhance the conclusions.
- Although the district is approximately twenty-two percent minority, the data reflected the aggregate regular education population. There was no disaggregation of minority data. Additional studies that disaggregate minority population(s) could provide data to enhance the conclusions.

References

Abbott et al. v Burke et al. 119 NJ287. (1997).

Alaya, A. (1999, February 7). As scores slide, teachers happy to write off old test. Sunday Star-Ledger, 1, 29.

Anastasi, A. (1973, 1959). Psychological testing. New York: The Macmillan Company.

Anastasi, A. (1982). Contributions to differential psychology: Selected papers. New York: Praeger Publishers.

Archer, J. (1999, March 17). R.I. halts exams in wake of wide-scale security breaches. Education Week, XVIII (27), 28.

Associated Press (AP). (2000, June 8). Educators cheat in pupil tests. Asbury Park Press, A8.

Bailey, K. (1987). Methods of social research, 3rd edition. New York: The Free Press.

Barry, J. S. and Hederman, R. S. (1998, December). Report card on american education. A state-by state analysis 1976-1998. Washington, D.C.: American Legislative Exchange Council.

Berman, I. (1999). Academic standards: An overview. in Focus on Education, 1999 (43). Jamesburg, NJ: New Jersey Association for Supervision and Curriculum Development.

Bezy, K. G. (1999, December). State standards fuel innovation and collaboration. The High School Magazine, 7 (4), 5-7.

Blair, J. (1999, December 15). Mismatched curricula leave freshmen ill-prepared, study finds. Education Week, XIX (16), 9.

Board of Directors, International Reading Association. (1999, November). High-stakes assessments in reading. A position statement of the International Reading Association. The Reading Teacher, 53 (3), 257-263.

Bock, R. D., Mislevy, R. and Woodson, C. (1982, March). The next stage in educational assessment. Educational Researcher, 11, 4-15.

Boser, U. (1999, June 23). Study finds mismatch between Calif. standards and assessments. Education Week, XVIII (41), 10.

Boser, U. (2000, June 21). Opponents of high-stakes tests seek to breach exam security. Education Week, XIX (41), 14-15.

Bowman, D. (2000, February 2). Test errors irk Vermont education dept. Education Week, XIX (21), 22.

Bracey, G. (1995, April). Portfolios in Vermont. Phi Delta KAPPAN, 646-647.

Bradley, A. (1997, February 19). Phila. assailed for large numbers of 11th grade 'no shows' on tests. Education Week, XVI (21), 7.

Bradley, A. (1999, April 21). Countersuit filed over test breach. Education Week, XVIII (32), 4.

Bradley, A. (2000, March 22). L.A. proposes linking teacher pay to tests. Education Week, XIX (28), 3.

Bradley, A., Hofff, D. and Manzo, K. K. (1999, October 27). Teachers support most standards-based changes. Education Week, XIX (9), 8.

Brandt, R. (1992). On performance assessment: A conversation with Grant Wiggins. Educational Leadership, 49, 35-37.

Burch, C.B. (1997, January). Creating a two-tiered portfolio rubric. English Journal.

Chase, B. (1999, March). Testing, testing. (Press release). Washington, DC: National Education Association

Chiles, N. (1998, September 27). Scholarship rule hurts Jersey kids. Sunday Star-Ledger, 1,9.

Cizek, G. J. (1998, October). Filling in the blanks: Putting standardized tests to the test. Fordham Report, 2 (11).

Cohen, R., Swerdlik, M., and Smith, D. (1992). Psychological testing and assessment, second edition. Mountain View, CA: Mayfield Publishing Company.

Cooperman, S. (1999, October 3). Educators scrambling to explain bleak student test scores to public. The Sunday Star-Ledger, section 10, p. 7.

Cronbach, L. J. (1984). Essentials of Psychological Testing, fourth edition. New York: Harper & Brothers.

CTB/McGraw-Hill. (1997). The only one: Terra nova. Monterey, CA:Author.

DeMonte, J. and Rapp, L. (1998, September). The top 75 public high schools. New Jersey Monthly, XXIII (9), 52-57.

Education Trust. (1999, December). Ticket to nowhere: The gap between leaving high school and entering college and high-performance jobs. Washington, DC: Education Trust.

Educational Testing Service. (1995). Comprehensive testing program III technical report. New Jersey: author.

Education Week. (1996, December 11). Firm corrects Maine test. XVI (15), 4.

Education Week. (1998a, May 20). News in brief. XVII (36), (23).

Eha, L. (ed.). (1998, Fall). Test item formats. The Evaluator. New York, NY: Educational Records Bureau.

Eha, L. (ed.). (2000, Summer). Study compares CTP III and a state testing program. The Evaluator. New York, NY: Educational Records Bureau.

Farr, R. (1990). Trends: Reading: Setting directions for language arts portfolios. Educational Leadership, 48, 103.

Feuer, M., & Fulton, K. (1993). The many faces of performance assessment. Phi Delta Kappan, 74, 478.

Finn, C. E., Jr. and Petrilli, M., eds. (2000, January). The state of the state standards 2000. Washington, D.C.: The Thomas B. Fordham Foundation.

Fullan, M. and Stiegelbauer, S. (1991). The new meaning of educational change. New York: Teachers College Press.

Gearhart, M., Herman, J., Baker, E. & Whittaker, A. (1993).Whose work is it? A question for the validity of large-scale portfolio assessment (CSE Tech. Rep. 363). Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing.

Gillespie, C. S., Ford, K. assessment: Some question, some answers, some recommendation. Journal of Adolescent & Adult Literacy, 39, 480-491.

Glasserman, P. (1999). Statistics. (Student notes). New York: Columbia Business School.

Glovin, D. (1998, November 29). Low-paid part-timers judge N.J. students. The Record Online: http://www.bergen.com/ed/testdg199811291.htm, 1-12.

Grace, C. (1992). The portfolio and its use: Developmentally appropriate assessment of young children.

Urbana, IL: Clearinghouse on Elementary and Early Childhood Education. (ERIC Document Reproduction Service No. ED 357 393)

Gronlund, N. E. (1981). Measurement and evaluation in teaching (fourth edition). New York: Macmillan Publishers Co, Inc.

Guion, R. D. (1976). Recruitment, selection, and job placement. In M. D. Dunnette (Ed.), Handbook of industrial and organizational psychology, 777-828. Chicago: Rand McNally.

Harris, S. (1999, February 16). School boards attempt to stop cheating on tests. The Times (Shreveport, Bossier City, Ark-La-Tex), 1, 2A.

Henry, T. (2000, January 10). Ruling opens door to more high school exit tests. USA TODAY, 6D.

Herman, J. & Winters, L. (1994, October). Portfolio research: A slim collection. Educational Leadership, 48-53.

Hespe, D. (1999a, May 5). Time line for review of proposed standards and assessment code. (memo: Members, State Board of Education). Trenton, NJ: New Jersey Department of Education.

Hespe, D. (1999b, May 21). 2000 National Assessment of Educational Progress (NAEP). (memo, Chief School Administrators). Trenton, NJ: New Jersey Department of Education.

Hespe, D. (1999c, September 20). ESPA/GEPA press release. (memo, Chief School Administrators). Trenton, NJ: New Jersey Department of Education.

Hespe, D. (1999d, December 30). Students need higher academic standards to succeed - ESPA, GEPA offers that. (Press Release). Trenton, NJ: New Jersey Department of Education.

Hespe, D. (2000a, January 10). State withdrawal from the February 2000 National Assessment of Educational Progress (NAEP) assessment. (memo: Superintendents declining to participate in the 2000 NAEP state assessment). Trenton, NJ: New Jersey Department of Education.

Hespe, D. (2000b, April 5). Standards and assessment for student achievement. (memo: Members, State Board of Education). Trenton, NJ: New Jersey Department of Education.

Hespe, D. (2000c, April 17). Revisions to the 2000-2001 ESPA and GEPA assessment schedule. Examining the optimal assessment of visual and performing arts, health and physical education, and world languages. (memo: Chief School Administrators). Trenton, NJ: New Jersey Department of Education.

Hespe, D. (2000d, August 21). 2000 elementary school proficiency assessment (ESPA) and grade eight proficiency assessment (GEPA) results. (Memo: Chief School Administrators,

Directors of Charter Schools). Trenton, NJ: New Jersey Department of Education.

Heubert, J. and Hauser, R., eds. (1999). High stakes testing for tracking, promotion, and graduation. Washington, D.C.: National Academy Press.

Heyboer, K. (1998, September 2). SAT results hold steady. The Star-Ledger, 19, 24.

Hills, T. W. (Chairperson). (1988). Integrating mathematics, reading, and writing instruction in kindergarten-first and second-third grades. (HSPT Institutes). Trenton, NJ: New Jersey Department of Education Division of General Academic Education.

Hirsch, E. D. Jr. (2000, February 2). The tests we need and why we don't quite have them. (Commentary). Education Week, XIX (21), 40-41, 64.

Hoff, D. (1997b, March 26). Chiefs' group backs Clinton testing proposals. Education Week, XVI (26), 19, 21.

Hoff, D. (1997c, June 25). Local Control could stymie Clinton Tests. Education Week, XVI (39), 1,26-27.

Hoff, D. (1997d, October 15). GOP plays hardball to block national tests. Education Week, XVII (7), 24, 26.

Hoff, D. (1997e, October 29). Reading bill makes progress; testing doesn't. Education Week, XVII (9), 19, 24.

Hoff, D. (1998a, September 9). Many problems continue to haunt Clinton's proposed national tests. Education Week, XVIII (1), 34, 39.

Hoff, D. (1998b, September 30). Panel assails assessment calculations. Education Week, XVIII (4), 1, 23.

Hoff, D. (1998c, October 14). National testing plan appears headed for perilous end. Education Week, XVIII (7), 22.

Hoff, D. (1998d, October 21). As tests get slight reprieve, governing board forges on. Education Week, XVIII (8), 22.

Hoff, D. (1998e, December 2). Research council pledges help in setting NAEP levels. Education Week, XVIII (14), 24.

Hoff, D. (1999a, February 3). Panel to probe validity of N.Y. reading test. Education Week, XVIII (21), 3.

Hoff, D. (1999b, June 16). Lessons of a century. The assessment culture. Education Week, XVIII (40), 20-27.

Hoff, D. (1999c, July 14). NAGB will keep achievement levels-for time being. Education Week, XVIII (42), 24.

Hoff, D. (1999d, October 13). Bush outlines broad testing plan for schools. Education Week, XIX (7), 24.

Hoff, D. (1999e, October 20). Mass. to rate schools based on state test scores. Education Week, XIX (8), 19.

Hoff, D. (1999f, December 1). Testing ETS. Education Week, XIX (14), 28-33.

Hoff, D. (1999g, December 15). N.Y.C. probe levels test-cheating charges. Education Week, XIX (16), 3.

Hoff, D. (2000a, January 12). States grades inch upward on content standards. Education Week, XIX (17), 5.

Hoff, D. (2000b, March 22). Testing foes hope to stroke middle-class ire. Education Week, XIX (28), 24, 31.

Hoff, D. (2000c, May 31). Massachusetts to put math teachers to the test. Education Week, XIX (38), 16, 18.

Hoff, D. (2000d, June 21). As stakes rise, definition of cheating blurs. Education Week, XIX (41), 1, 14-16.

Jacobson, L. (1997, May 7). Kentucky. Education Week, XVI (32), 7.

Jacobson, L. (1999, February 17). District questions fairness of accountability proposals. Education Week, XVIII (23), 5.

Jerald, C. D. (2000, January 13). The state of the states. Education Week (Quality Counts), XIX (18), 62-65.

Joftus, S. and Berman, I. (1998, January). Great expectations? Defining and assessing rigor in state standards for mathematics and English language arts. (Special report). Washington, D.C.: Council for Basic Education.

Johns, J. (1992). How professionals view portfolio assessment. Reading Research and Instruction, 32, 1-10.

Johnson, C. (1985, September 16). HSPT material. (memo: Robert Osak, Paul Winkler). Trenton, NJ: Department of Education.

Johnston, R. (1998, April 29). In Texas, the arrival of spring means the focus is on testing. Education Week, XVII, (33), 1, 20-21.

Johnston, R. (1999, March 17). Texas presses districts in alleged test-tampering cases. Education Week, XVIII (27), 22, 28.

Johnston, R. and M. Galley. (1999, April 14). Austin district charged with test tampering. Education Week, XVIII, (31), 3.

Johnston, R. and L. Jacobson. (1999, July 14). Testing fate. Education Week, XVIII (42), 17.

Kearns, J. F., Kleinert, H. L. and Kennedy, S. (1999, March). We need not exclude anyone. Educational Leadership, 56 (6), 33-38.

Kirst, M. (1998, September 9). Bridging the remediation gap. Education Week, XVIII (1), 52, 76.

Klagnolz, L. (1997, December 5). Policy paper on standards and assessment for student achievement. (memo: Colleague). Trenton, NJ: Department of Education.

Klagholz, L. (1998). Standards and assessment for student achievement. (memo: Members, State Board of Education). [WWW document] URL http://www.state.nj.us./njded/proposed/standards/stass2.htm.

Koretz, D. (1993). New report of the Vermont Portfolio Project documents challenges. National Council on Measurement in Education Quarterly Newsletter, 1 (4), 1-2.

Koretz, D., Stetcher, B., Klein, S. & McCaffrey, D. (1994). The Vermont Portfolio Assessment Program: Findings and implications. Educational Measurement: Issues and Practice, 13 (3).

Lawton, M. (1996, October 9). PSAT to add writing test to settle bias case. Education Week, XVI (6), 9.

Lawton, M. (1996b, November 6). Survey seeks to identify 'essential' standards. Education Week, XVI (10), 5.

Lawton, M. (1996c, November 13). Alleged tampering underscores pitfalls of testing. Education Week, XVI (11), 5.

Lawton, M. (1997a, April 2). Cost of test proposal up to $10 million in first year. Education Week, XVI (27), 21.

Lawton, M. (1997b, June 25). Wilson proposal for basic-skills test in Calif. draws fire. Education Week, XVI (39), 14.

Lawton, M. (1997c, October 1). Riley delays national tests' development. Education Week, XVII (5), 1, 26.

Lawton, M. (1997d, October 22). Discrimination claimed in Texas exit-exam lawsuit. Education Week, XVII (8), 3.

LeGlise, S. (1999, June 9). Letter to Commissioner Hespe from NJASA. Trenton, NJ: New Jersey Association of School Administrators.

LeMahieu, P., Gitomer, D. and Eresh, J. (1995b, Fall). Portfolios in large-scale assessment: Difficult but not impossible. Educational Measurement: Issues and Practice, 11-28.

Lemann, N. (1999). The big test. New York: Farrar, Straus and Giroux.

Levine, D.; Berenson, M.; and Stephan, D. (1999). Statistics for managers using Microsoft excel, second edition. Upper Saddle River, NJ: Prentice-Hall, Inc.

Lindsay, D. (2000, April 5). Contest. Education Week, XIX (30), 30-37.

Manzo, K. (1996a, October 30). Phila. plan links student achievement, teacher pay. Education Week, XVI (9), 3.

Manzo, K. (1996b, December 4). Vt. to combine standardized tests with portfolios. Education Week, XVI (14), 3.

Manzo, K. (1997, October 22). High stakes: Test truths or consequences. Education Week, XVII (8), 1,9.

Manzo, K. (1998a, January 28). For girls, writings on the wall in new PSAT exam. Education Week, XVII (20), 3.

Manzo, K. (1998b, September 9). Report for goals panel calls for consensus on standards. Education Week, XVIII (1), 8.

Manzo, K. (1999a, June 9). Group to launch new international assessment. Education Week, XVIII (39), 5.

Manzo, K. (1999b, October 6). U.S. students lack writing proficiency. Education Week, XIX (6), 1,18.

Manzo, K. (1999c, November 10). Guidelines on student assessment released. Education Week, XIX (11), 12.

Manzo, K. (2000, March 15). NAEP drops long-term writing data. Education Week, XIX (27), 1, 17.

McGettigan, K. (1989, June). Report of the eleventh-grade high school proficiency test reading skills development committee. (PTM 900.37). Trenton, NJ: New Jersey State Department of Education.

McGettigan, K. (1990, May). Report of the reading committee: Identification of the 8th-grade skills in reading and test specifications and sample items for the 11th-grade high school proficiency test and the 8th-grade early-warning test.

(PTM 1000.47). Trenton, NJ: New Jersey State Department of Education.

Messick, S. (1993). Validity. In R. Linn (Ed.), Educational measurement, third edition, 13-103. Phoenix, AZ: American Council on Education and The Oryx Press.

Micklo, S. (1997, Spring). Math portfolios in the primary grades. Childhood Education, 194-199.

Millman, J. and Greene, J. (1993). The specifications and development of tests of achievement and ability. In R. Linn (Ed.), Educational measurement, third edition, 335-366. Phoenix, AZ: American Council on Education and The Oryx Press.

Mooney, J. (2000a, April 6). State standards adopted, minus a few radical ideas. The Star-Ledger, 18.

Mooney, J. (2000b, April 12). N.J. hires firm to grade its school standards and tests. The Star-Ledger, 25.

Mooney, J. (2000c, August 23). Fault lies in the 4th-graders' test. The Star-Ledger, 1, 26, 27.

Moss, P. and Schutz, A. (1999, May). Risking frankness in educational assessment. KAPPAN, 80 (9), 680-87.

Mursell, J.L. (1947). Psychological testing. New York: Longmans, Green and Co.

National Computer Systems. (2000, April 20). Errata. (FAX). Author.

New Jersey Association of School Administrators. (2000, September 14). Executive committee motions regarding state assessments. Trenton, NJ: author.

New Jersey Department of Education. (1985, September). High School Proficiency Test skill array: Reading. (PTM 400.94). Trenton, NJ: author.

New Jersey Department of Education. (1990, May). Report of the mathematics committee: Identification of the 8th-grade skills in mathematics and test specifications and sample items for the 11th-grade High School Proficiency Test and the 8th-grade Early Warning Test. (PTM 1000.48). Trenton, NJ: author.

New Jersey Department of Education. (1993, July). Summary of standardized test results for grades three and six. Trenton, NJ: author.

New Jersey Department of Education. (1995a, March). Cycle I school and district guidelines: How to interpret and use EWT reports. (PTM 1346.00). Trenton, NJ: author.

New Jersey Department of Education. (1995b, November). Comprehensive plan for educational improvement and financing. Trenton, NJ: author.

New Jersey Department of Education. (1997a). Grade 8 Early Warning Test (EWT) parent information. (PTM 1400.43). Trenton, NJ: author.

New Jersey Department of Education. (1997b, March). Grade 8 Early Warning Test March 1997 district and school test coordinators' manual. (PTM 1400.46). Trenton, NJ: author.

New Jersey Department of Education. (1997c, March). Examiner's manual grade 8 Early Warning Test March 1997. (PTM 1400.47). Trenton, NJ: author.

New Jersey Department of Education. (1997d, June). March 1997 Grade 8 Early Warning Test (EWT) cycle I school and district guideline: How to interpret and use the EWT reports. (PTM 1400.74). Trenton, NJ: author.

New Jersey Department of Education. (1997e, September). March 1997 Grade 8 Early Warning Test (EWT) cycle II the registered holistic scoring method: a writing handbook. (PTM 1400.92). Trenton, NJ: author.

New Jersey Department of Education. (1998, June). March 1998 Grade 8 Early Warning Test (EWT) cycle 1 school and district guidelines: How to interpret and use EWT reports. (PTM 1500.35). Trenton, NJ: author.

New Jersey Department of Education. (1999a). Your guide to the HSPT11. (PTM 1501.21). Trenton, NJ: author.

New Jersey Department of Education. (1999b, August). School and district guidelines: Interpretation and use of individual

student reports and rosters for GEPA and ESPA. (PTM 1501.24). Trenton, NJ: author.

New Jersey Department of Education. (1999c). A parent's guide to the new grade eight proficiency assessment. Trenton, NJ: author.

New Jersey Department of Education. (1999d). A parent's guide to the new elementary school proficiency assessment. Trenton, NJ: author.

New Jersey Department of Education. (1999e, December). 1997 cohort state summary. Trenton, NJ: author.

New Jersey Department of Education. (1999f, December). May 1999 elementary school proficiency assessment. State summary. Trenton, NJ: author.

New Jersey Department of Education. (1999g, December). March 1999 grade eight proficiency assessment. State summary. Trenton, NJ: author.

New Jersey Department of Education. (2000a, January). Cycle I school and district guidelines: How to interpret HSPT11 reports. Trenton, NJ: author.

New Jersey Department of Education. (2000b, February). Cycle II school and district guidelines: How to interpret and use ESPA and GEPA school and district reports. (PTM 1501.59). Trenton, NJ: author.

New Jersey Department of Education. (2000, April). New Jersey administrative code title 6A standards and assessment for student achievement. (Chapter*[6] 8*). Trenton, NJ: author.

New Jersey Principals and Supervisors Association (1998a, Spring/Summer). Final edict in Abbot v. Burke? NEWSLETTER (Capitol Update), 1,3-4.

New Jersey Principals and Supervisors Association (1998b, Spring/Summer). New funding model proposed. NEWSLETTER (Capitol Update), 1, 5.

New Jersey Principals and Supervisors Association (1999, December). Delay urged in negative consequences of state testing. NEWSLETTER, 1.

Nitko, A. (1983). Educational tests and measurement. An introduction. New York: Harcourt Brace Jovanovich, Inc.

Nusser, N. and Faris, D. (2000, September). The top 75 public high schools. New Jersey Monthly, 25 (9), 65-69.

Olson, L. (1997, February 19). Focus on basics key to Clinton call for testing. Education Week, XVI (21), 1, 19.

Olson, L. (2000a, February 16). New service will help compare district's spending with results. Education Week, XIX (23), 5.

Olson, L. (2000b, April 5). Worries of a standards 'backlash' grow. Education Week, XIX (30), 1, 12-13.

Olson, L. and Hoff, D. (1999, October 6). Teaching tops agenda at summit. Education Week, XIX (6), 1,20.

Patton, P. and Thompson, T. (1999, October 13) Continuity of purpose and a common vocabulary. Education Week, XIX (7), 52.

Peretzman, P. (1996, July 10). Curriculum standards to become central focus for state board policies and regulations. (Press release). Trenton, NJ: New Jersey State Department of Education.

Petersen, N., Kolen, M. and Hoover, H. (1989). Scaling, norming, and equating. In R. Linn (Ed.), Educational measurement third edition, 221-262. New York: American Council on Education and Macmillan Publishing Company.

Popham, W. J. (1975). Educational evaluation. Englewood Cliffs, NJ: Prentice-Hall, Inc.

Popham, W. J. (1981). Modern educational measurement. Englewood Cliffs, NJ: Prentice-Hall.

Popham, W. J. (1999, May 12). Assessment apathy (Commentary). Education Week, XVIII (35), 32.

Popham, W. J. and Lindheim, E. (1981). Implications of a landmark ruling on Florida's minimum competency test. Phi Delta Kappan, 63 (1), 18-20.

Portner, J. (1999, February 17). Most schools failed, but experts call Va. tests fair. Education Week, XVIII (23), 3.

Public Agenda (2000, February 16). Reality check 2000. Education Week, XIX (23), S1-S8.

Rimbach, J. and Wiggins, O. (2000, August 23). 4th-grade language arts scores deemed poor, prompting review. http://www.bergen.com/ed/testingjr20000823.htm.

Robelen, E. (1999, October 6). Lawmakers debate accountability's meaning. Education Week, XIX (6), 26, 31.

Robelen, E. (2000, May 24). La. Set to retain 4th, 8th graders based on state exams. Education Week, XIX (37), 24.

Rosenholtz, S. J. (1991). Teacher's workplace: The social organization of schools. New York: Teachers College Press.

Rothstein, R. (1999, December 8). In judging schools, one standard doesn't fit all. www.nytimes.com/library/national, author.

Sack, J. (2000, March 1). Riley urges "review" of standards. Education Week, XIX (25), 1, 32.

Sandham, J. (1999a, April 7). Exam-testing breaches put focus on security. Education Week, XVIII (30), 20

Sandham, J. (1999b, July 14). In first for states, Florida releases graded 'report cards' for schools. Education Week, XVIII (42), 18.

Sandham, J. (2000, April 5). Colorado lawmakers OK school rating plan. Education Week, XIX (30), 20.

Schechter, E. (1997, October 16). New Jersey's core curriculum content standards and related issues. (memo). Trenton, NJ: New Jersey Department of Education.

Schechter, E. (1998, March 20). Assessment update. (memo: Chief School Administrators). Trenton, NJ: New Jersey Department of Education.

Schechter, E. (1999, October 22). Releasing additional 1999 ESPA and GEPA results. (memo: Chief School Administrators). Trenton, NJ: New Jersey Department of Education.

Schmoker, M. and Marzano, R. (1999, March). Realizing the promise of standards-based education. Educational Leadership, 56 (6), 17-21.

Schwartz, T. (1999, January 10). Is this any way to run a meritocracy? The S.A.T. numbers game. The New York Times Magazine, Section 6, 30-35, 51, 56.

Sergiovanni, T. (2000, February 16). Changing education change. Education Week, XIX (23), 27, 31.

Smith, F. (1986). Insult to intelligence. New York: Arbor House.

Stake, R. (1998). Some comments on assessment in U.S. education. Education Policy Analysis Archives (on-line serial), 6 (14). hhtp://epaa.asu.edu/epaa/ von14.html.

Stotsky, S. (1997, July). State English standards. Fordham Report, 1 (1).

Thorndike, E.L. (1906). The principles of teaching. New York: A.G.Seiler.

Tirozzi, G. (1998, August 5). It's about teaching and learning-not testing. Education Week, XVII (43), 44,47.

Viadero, D. (1999a, October 6). Stanford report questions accuracy of tests. Education Week, XIX (6), 3.

Viadero, D. (1999b, October 20). CTB knew of problems earlier, Indiana districts say. Education Week, XIX (8), 3.

Viadero, D. (1999c, November 3). Wrong forms used for scores. Education Week, XIX (10), 4.

Viadero, D. and Blair, J. (1999, September 29). Error affects test results in six states, Wrong forms used for scores. Education Week, XIX (5), 1, 13-15.

White, K. (1999, June 2). Student protesters in Massachusetts sit out state exams. Education Week, XVIII (38), 14,15.

Wiggins, G. (1994). CLASS slide presentation, 39.

Witte, R. (1993). Statistics fourth edition. New York: Harcourt Brace Jovanovich College Publishers.

Bibliography

Campbell, D. (1967). Recommendations for APA test standards regarding construct, trait, or discriminant validity. In D.N. Jackson and S: Messick (eds.), Problems in human assessment. New York, McGraw Hill.

Campbell, D. (2000, January). Authentic assessment and authentic standards. KAPPAN, 81 (5), 405-407.

Cannell, J. J. (1988). Nationally normed elementary achievement testing in America's public schools: How all 50 states are above average. Educational Measurement: Issues and practice, 7 (2), 5-9.

Cronbach, L. J. & Meehl, P. E. (1955). Construct validity in psychological tests. Psychological Bulletin, 52, 281-302.

Cronbach, L. J. (1960). Essentials of Psychological Testing, second edition. New York: Harper & Brothers.

Cronbach, L. J. (1980). Validity on parole: How can we go straight? in Measuring achievement: Progress over a decade. San Francisco: Jossey-Bass.

Cronbach, L. J. (1988). Five perspectives on the validity argument. in Test Validity (H. Wainer, ed.). Hillsdale, NJ: Erlbaum.

Cronbach, L. J.; Gleser, G. C; Nanda, H; Nageswari, R. (1972). The dependability of behavioral measurements: Theory of

generalizability for scores and profiles. New York: John Wiley & Sons, Inc.

Cross, C. (1998, October 21). The standards wars: some lessons learned. Education Week, XVIII (8), 32,35.

Educational Research Service (ERS). (1999). Explaining testing and test scores to parents. The Informed Educator Series. Arlington, VA: author.

Eisner, E. (1999, May). The uses and limits of performance assessment. KAPPAN, 80 (9), 658-61.

Farr, R. (1970). Measurement and evaluation of reading. New York: Harcourt, Brace & World, Inc.

Farr, R., & Tone, B. (1994). Portfolio and performance assessment. New York: Harcourt Brace.

Feldt, L. & Brennan, R. (1989). Reliability in Educational measurement, 3rd edition. New York: Macmillan.

Galley, M. (1999, March 24). Principal resigns over test leak. Education Week, XVIII (28), 4.

Glaser, R. and Linn, R. (1993). Foreword. In L. Shepard (Ed.), Setting performance standards for student achievement. Stanford, CA: National Academy of Education, Stanford University.

Haertel, E. (1999, May). Performance assessment and education reform. KAPPAN, 80 (9), 662-67.

Hendrie, C. (1997, June 11). Tougher tests spur debate on N.Y. diploma. Education Week, XVI (37), 9.

Hendrie, C. (1999, June 2). Poor districts fare worst on N.Y. assessment. Education Week, XVIII (38), 14.

Hess, F. and Brigham, F. (2000, January). The promise and peril of high-stakes testing. American School Board Journal, 187 (1), 26-29.

Hoff, D. (1998, October 21). Kentucky to include norm-referenced test in accountability plan. Education Week, XVIII (8), 16.

Hoff, D. and A. Coles. (1999, March 24). Security Breach. Education Week, XVIII (28), 17.

Hurwitz, N. and Hurwitz, S. (2000, January). Do high-stakes assessments really improve learning? American School Board Journal, 187 (1), 20-25.

Johnston, R. (1997a, May 28). Dispute over KY test section sparks broader debate. Education Week, XVI (35), 12.

Johnston, R. (1997b, September 24). At session's close, Calif. lawmakers ok test plan. Education Week, XVII (4), 17.

Johnston, R. (1997c, October 22). Mich. house passes bill to revise high school testing program. Education Week, XVII (8), 11.

Johnston, R. (1999a, March 24). Assessment. Education Week, XVIII (28), 8.

Johnston, R. (1999b, March 31). Reform bills pass in Calif. Legislature. Education Week, XVIII (29), 1, 18.

Johnston, R. (1999c, July 14). Education spotlighted in California's fiscal 2000 budget. Education Week, XVIII (42), 17.

Kish, C.K., Sheehan, J.K., Cole, K.B., Struyk, L.R. and Dinder, D. (1997, April-May). Portfolios in the classroom: A vehicle for developing reflective thinking. The High School Journal, 254-260.

Lawton, M. (1997, March 19). Testing ventures tied to standards take flight. Education Week, XVI (25), 9, 18.

LeMahieu, P., Gitomer, D. and Eresh, J. (1995a). Portfolios beyond the classroom: Data quality and qualities. (Center for Performance Assessment Report No. MS-94-01.) Princeton, NJ: Educational Testing Service.

Lindsay, D. (1996, October 2). Whodunit? Someone cheated on standardized tests at a Connecticut school. And it wasn't the students. Education Week, XVI (5), 25-29.

Madaus, G. and O'Dwyer, L. (1999, May). A short history of performance assessment: Lessons learned. KAPPAN, 80 (9), 688-95.

Maeroff, G. (1991). Assessing alternative assessment. Phi Delta Kappan, 73, 272-281.

Malhotra, Y. (1993). An analogy to a competitive intelligence program: Role of measurement in organizational research [WWW document]. URL http://www.brint.com/papers/compint.htm.

Moje, E., Brozo, W. and Hass, J. (1994). Portfolios in a high school classroom: Challenges to change. Reading Research and Instruction, 33, 275-292.

Myers, J. (1991). How that literacy happens in contexts, how do we know if the contexts are authentic? In J. Zutell & S. McCormick (Eds.), Learner factors/teacher factors: Issues in literacy research and instruction, 91-96. Chicago: National Reading Conference.

Myford, C. M., & Mislevy, R. J. (1995). Monitoring and improving a portfolio assessment system. (Center for Performance Assessment Report No. MS-94-05.). Princeton, NJ: Educational Testing Service.

National Education Goals panel. (1998). Talking about tests: An idea book for state leaders. Washington, DC: U.S. Government Printing office.

Natt, J. SAT/ACT scores, test-takers increase during the 1990s. Leadership News, 2 (3), 6.

New Jersey Department of Education. (1991, August). Guide to procedures for scoring the reading constructed-response items. (PTM 1059.00). Trenton, NJ: author.

New Jersey Department of Education. (1995c, December). State summary. Trenton, NJ: author.

Ohanian, S. (2000, January). Goals 2000: What's in a name. KAPPAN, 81 (5), 344-355.

O'Neil, J. (1999, March). Core knowledge & standards, A conversation with E. D. Hirsch Jr. Educational Leadership, 56 (6), 28-31.

Phillips, G. et al. (1993, April). Interpreting NAEP scales. Washington, DC: OERI.

Pickett, W. and Burrill, D. (1994, September). The use of quantitative evidence in research: A comparative study of two literatures. Educational Researcher, 23 (6), 18-21.

Popham, W. J. (1999, March). Why standardized tests don't measure educational quality. Educational Leadership, 56 (6), 8-15.

Popkewitz, T. S. (1984). Paradigm and ideology in educational research. London: Falmer.

Ramirez, A. (1999, November). Assessment-driven reform: The emperor still has no clothes. Phi Delta KAPPAN, 81 (3), 204-208.

Ravitch, D. (1998, December 16). What if research really mattered (Commentary). Education Week, XVIII (16), 33,34.

Sirotnik, A. and Kimball, K. (1999, November). Standards for standards-based accountability systems. Phi Delta KAPPAN, 81 (3), 209-214.

Supovitz, J. and Brennan, R. (1998). Mirror, mirror, on the wall, which is the fairest test of all? An examination of the equitability of portfolio assessment relative to standardized tests. Cool Thinking on Hot Topics. Harvard Educational Review.

The College Board. (1998). Working with the PSAT/NMSQT. New York: author.

The College Board. (1999). How to understand/interpret/explain/use SAT program services/scores/publications/software/tests. New York: author.

The Princeton Review. (1997). The new PSAT. www.review.com: author.

Walker, H. M. (1943). Elementary statistical methods. New York: Henry Holt and Co.

Wiggins, G. (1993). Assessing student performance. San Francisco, CA: Josey-Bass Publishers.

Wiggins, G. (1998). Educative assessment. Designing assessments to inform and improve student performance. San Francisco, CA: Josey-Bass Publishers.